Advanced diffusion MRI analysis methods for neonatal imaging

Pietsch, Maximilian Rainer

Awarding institution:
King's College London

ADVANCED DIFFUSION MRI ANALYSIS METHODS FOR NEONATAL IMAGING

Maximilian Pietsch

A thesis submitted for the degree of Doctor of Philosophy

April 2018

Contents

Abbreviations

ADC Apparent Diffusion Coefficient. 53, 217

AFD Apparent Fibre Density. 59

AP average precision. 81–84, 87, 88, 91, 294

AUROC area under the receiver operator curve. 81–85, 87, 92, 95

CC corpus callosum. 28, 203, 206, 218, 225, 227, 230, 297

CNN Convolutional Neural Network. 64, 70, 71

CP cortical plate. 16, 21, 22, 26, 29

CSD constrained spherical deconvolution. 59, 237

CSF cerebrospinal fluid. 19, 30, 60, 61, 193, 196, 203, 204, 213, 217–223, 230, 237, 296

CST corticospinal tract. 35, 206, 209, 214, 225, 232, 233, 296, 297

dHCP Developing Human Connectome Project. 10, 11, 115, 152, 203, 206, 216, 217, 224, 234, 236, 237, 296

dMRI diffusion weighted MRI. 10, 61, 62, 98, 99, 102, 191

DSI Diffusion Spectrum Imaging. 53

DTI Diffusion Tensor Imaging. 10, 53, 187, 207

DW diffusion weighted. 224, 225, 230, 232

EPI Echo Planar Imaging. 98

FA Fractional Anisotropy. 55, 58, 170, 193, 219, 230, 231, 235, 297

FOD Fibre Orientation Distribution. 58–60

GM grey matter. 60, 61, 193, 196–205, 207, 213, 218, 219, 221, 227, 230–232, 296, 297

HARDI High Angular Resolution Diffusion Imaging. 10, 55, 58, 60, 61, 115, 187, 191, 192, 196, 202, 203, 206, 213, 216, 236, 237

HCP Human Connectome Project. 196, 198, 199, 211, 212, 218, 296

IZ intermediate zone. 21, 22, 24, 26, 29

MBP myelin basic protein. 33, 34

MCC Matthews correlation coefficient. 80, 81, 84, 85, 87, 92

MSMT-CSD multi-shell multi-tissue constrained spherical deconvolution. 61, 62, 196, 202, 204, 216–218, 223, 224, 232

MZ marginal zone. 21–24

ODF orientation distribution function. 61, 194, 196, 199, 201, 204, 205, 207, 212, 213, 224, 296

ODFs orientation distribution functions. 61, 193, 195, 196, 199–202, 206–212, 218, 228, 296

PMA postmenstrual age. 218, 222, 223, 225, 227–231, 233, 296

SP subplate. 16, 21, 22, 26, 29

SVZ subventricular zone. 16, 21, 24, 29

VZ ventricular zone. 16, 21, 26, 31

WM white matter. 61, 187, 193, 196–205, 207, 212, 213, 216–225, 230–232, 296, 297

Acknowledgements

It is easy to name a number of people who supported and helped me in the past months, making this thesis a reality, but it is hard to do their contributions justice. It was a challenging and interesting journey, on which I would have made many more detours and would have run out of steam without your support. First and foremost, I would like to thank my supervisor J-Donald Tournier. Donald, thank you for your excellent support and guidance, an always open door, and for the freedom to learn and explore what interested me the most. Thank you for the hours you spent commenting on this thesis.

I would also like to especially thank Jonathan O'Muircheartaigh for his impressively fast and helpful feedback on this thesis. I have had a great time in the centre for the developing brain, where I found a supportive and collegial environment and a few friends.

Finally, I would like to thank my friends and family who kept me physically and mentally in shape, supported me financially, and condoned me wearing blinkers for a few months. I would like to dedicate this thesis to Heidi Pietsch.

Thank you, everyone! Grazie mille! Vielen Dank!

Maximilian Pietsch
26 April 2018

Chapter 1

Introduction

Developmental processes taking place during the third trimester and the neonatal period lay the foundation for a functioning human brain. In the course of these months, neuronal migration, cellular organisation, cortical development and myelination shape the form and function of our arguably most complex and outstanding organ.

Diffusion weighted MRI (dMRI) has been extensively used to study the rapid changes in microstructural properties of white and grey matter non-invasively and provides contrast that is complementary to other imaging modalities [Yoshida et al., 2013]. The sensitivity to processes on the cellular level has made diffusion imaging a tool for studying white matter development and the early detection of injury [Hüppi, Dubois, 2006].

Linking the measured signal to changes in the cellular composition and organisation of brain tissue poses data processing challenges unique to the pediatric population. In particular, movement during the acquisition corrupts diffusion images beyond repair and requires manual data cleaning. We developed a neural network classifier that can perform this task automatically, allowing large-scale automated processing and analysis of diffusion data.

Also, inferring cellular tissue properties from the signal is difficult as the brain simultaneously undergoes a number processes that could alter the contrast in various ways. In simulations, I investigate the validity of often implicitly assumed relations between quantities derived from Diffusion Tensor Imaging (DTI) and myelination in the context of changing tissue compartment volume fractions, showing that the interpretation of DTI parameters is flawed in the absence of a-priori knowledge about tissue microstructure.

In recent years, progress in acquisition and reconstruction techniques have facilitated acquiring quantitatively and qualitatively richer diffusion images. Currently, High Angular Resolution Diffusion Imaging (HARDI) and higher order diffusion models are uniquely positioned to capture and characterise developmental and maturation processes. The Developing Human Connectome Project (dHCP) is a group effort to advance the field of pediatric MRI and has made possible much of the work in this thesis. The HARDI data acquired as part of the dHCP captures microstructural properties of the developing brain with an unprecedented quality and information content.

Characterising tissue properties requires a model that allows inferring processes on the cellular level from HARDI data. To build this model, it is necessary to incorporate domain knowledge about physical and biological properties of brain tissue. Even for

adult populations, where brain tissue properties are comparatively static, developing higher order diffusion models that provide microstructure-specific markers is an open research question [Novikov, Kiselev, Jespersen, 2018]. For these reasons, this thesis investigates the use of data-driven techniques for the study of brain development, which do not require explicit a priori models of tissue microstructure, but rather attempt to decompose the observed signal into interpretable components.

In chapter 8, we develop tools to produce an unbiased group template of tissue properties at term, using a method that makes few assumptions about the microstructual properties of neonatal brain tissue. However, rapid brain maturation entails changes in tissue properties that require taking the temporal component into account. This term-time template is extended to the longitudinal domain in chapter 9, capturing tissue maturation patterns from 33 to 44 weeks gestational age in the dHCP cohort.

Together, these developments pave the way for detailed investigations into the development of the human brain. These techniques will form the basis for more advanced analyses, and will hopefully provide useful insights not available using existing methods.

Parts of this thesis and work related to experiments performed in this thesis have been presented at conferences under the titles "Effect of demyelination on diffusion tensor indices: A Monte Carlo simulation study" [Pietsch, Tournier, 2015], "Multi-contrast diffeomorphic non-linear registration of orientation density functions" [Pietsch et al., 2017a], "Transfer learning and convolutional neural net fusion for motion artefact detection" [Kelly et al., 2017], "Multi-shell neonatal brain HARDI template" [Pietsch et al., 2017b], and "Longitudinal multi-component HARDI atlas of neonatal white matter" [Pietsch et al., 2018].

A manuscript with the title "A framework for multi-component analysis of diffusion MRI data over the neonatal period" based on chapter 9 is currently under review in NeuroImage.

Part I.

Background

Chapter 2

Normal brain maturation in the perinatal period

Contents

2.1. Introduction

Understanding the developing human brain, be it for clinical or purely academic reasons, requires a temporal perspective on maturation processes. Most processes responsible for the structure of a functioning human brain start in the early embryonic weeks up to birth at around 40 weeks but the brain continues to mature at least into the second decade of life. This chapter summarises developmental processes that take place during the third trimester (27 weeks post conception until birth) and the neonatal period (birth to four postnatal weeks) with a focus on structural changes that are potentially detectable within this perinatal period, using diffusion MRI.

In this period, the morphology of the brain is relatively mature but the brain tissue undergoes rapid structural and functional development, which manifests in changes in

the shape, size, density, and arrangement of cells. Histological work conducted a century ago on the shape and arrangement of cells in the nervous system has led to remarkable progress in the understanding of how the brain functions [Llinás, 2003]. Following these seminal works, technical advancements such as immunofluorescence or electron microscopy allowed probing the constituents of brain tissue to a near-molecular level, which facilitated much more detailed and specific observations and led to new and refined understanding of the brain and the functions of its constituents [Sotelo, 2011].

In recent years, MRI has opened the possibility of measuring brain tissue properties non-invasively, which has led to numerous in-vivo studies of the normal, healthy human brain and its maturation. Yet, imaging newborns in a noisy MRI machine with lengthy sequences is challenging in many ways. Furthermore, interpreting the MRI signals, which integrate the tissue properties over at least a cubic millimetre and linking it to cellular properties is an ongoing field of research.

The aim of this chapter is to give a biological context to the signal changes we observe in the neonatal period with diffusion MRI. Therefore, I will not go into embryonic developmental processes such as the formation of the neural tube and the prosencephalon, or into the genetic or biochemical factors that orchestrate brain development but focus on neuronal migration and organisation, and on cortical development and myelination, which take place predominantly in the fetal and neonatal period.

2.2. Normal brain development

2.2.1. The big picture and timing of developmental events

The time before birth (prenatal period) can be split into two periods: the embryonic period, which consists of the first eight weeks post-fertilisation, which is followed by the fetal period, which lasts until birth. The neonatal period is the time from birth to one postnatal month.

In the embryonic period, features of the brain develop rapidly and the brain undergoes systematic morphological changes [O'Rahilly, Müller, 2010]. The first morphologic features of the nervous system appear as early as 3 weeks post-fertilisation. In the following fetal period, the brain develops by refining existing structures in a spatially and temporally varied manner, which is hard to characterise and subdivide into distinct morphological stages [O'Rahilly, Müller, 2006, chapter 4].

In the last fetal months and the neonatal period, the brain weight increases rapidly but, while the total number of cells increases, the density of DNA in the brain decreases in most parts of the brain [Dobbing, Sands, 1973]. The overall water density of brain tissue falls complementary to the increase in lipid concentration [Dobbing, Sands, 1973]. In this period, the growth of the brain is dominated by increasing cell size, branching and myelination [Dobbing, Sands, 1973] with the exception of the cerebellum whose cells are and remain comparatively small, leading to a steady increase in cell density (measured as increased DNA density [Dobbing, Sands, 1973]). In the neocortex, the total number of neurons remains constant between birth and 3 years of age but the number of glial cells (oligodendrocytes and astrocytes) increases linearly in that time span [Kjær et al.,

2016]. Note that there is controversy in the total numbers of neurons and glial cells and reported total numbers depend heavily on the cell counting methods used [Bartheld, Bahney, Herculano-Houzel, 2016].

The major developmental events leading to and during the neonatal period are summarised in figure fig. 2.1. The timeline serves as a frame of reference for the following sections but it can not depict the complex spatial heterogeneity of onset, peak occurrence and termination of developmental processes.

2.2.1.1. Gestational age

Unfortunately, multiple time scales have been defined to describe the age of a developing human, which have different origins and are typically used in different contexts [Engle, 2004] or confused with each other [O'Rahilly, Müller, 2000]. The term "gestational age" is commonly used to describe the age in the prenatal period but it is used differently in embryology and obstetrics and therefore discouraged by some researchers as being ambiguous [O'Rahilly, Müller, 2000]. In obstetrics, gestational age commonly refers to the time since "the first day of the last menstrual period" [Engle, 2004]. Note that gestational age is only defined up to birth after which the term postmenstrual age is used to refer to the gestational age plus the time since birth. A normal gestational age is defined as birth at 38 to 42 weeks gestational age [Engle, 2004].

However, the last menstrual period occurs about two weeks before ovulation. Hence, using the obstetric definition of gestational age, an embryo with gestational age of between zero and two weeks is yet to be conceived [O'Rahilly, Müller, 2006, chapter 5] making age a misnomer. It also introduces some biological variability that can span several days [Engle, 2004], which renders it a less desirable measure for embryonic time measure.

Therefore, in embryology, gestational age usually refers to the time since fertilisation. However, in practice, embryonic age, especially across species, is best characterised by morphometric measures [O'Rahilly, Müller, 2000].

However, defining a single timescale is undoubtedly useful. Since the dHCP defines the age of the babies at scan using the definition used in obstetrics, I will use this definition whenever I refer to the age or use the term "gestational age". Exceptions are descriptions of early embryonic events for which I explicitly state the age in weeks post-fertilisation.

2.2.1.2. Brain development in the third trimester and the neonatal period in a nutshell

Brain development in the weeks leading up to and post-birth, are governed by several developmental processes that affect the cortical morphology, which can be observed as a marked increase in gyrification (the number of many cerebral fissures, sulci and gyri) [Chi, Dooling, Gilles, 1977; Striedter, Srinivasan, Monuki, 2015]. Maturation on smaller length scales changes microstructural properties such as the cortical cytoarchitecture and the increased formation of connections in the brain.

In the third trimester, the cortex matures into its final six distinct horizontal layered

structure and the neurogenesis of new types of neurons starts [Honig, Herrmann, Shatz, 1996].

Neuronal migration [Sidman, Rakic, 1973; Kriegstein, Noctor, 2004] refers to the movement of neurons from their birthplace, usually the ventricular zone (VZ) or subventricular zone (SVZ), to their final destinations. Neuronal migration to the cerebral cortex has mostly finished before the third trimester [Rakic, 1990; Volpe, 2008] but at 28 weeks afferent neurons migrate from the subplate zone (subplate (SP)) to the cortical plate (cortical plate (CP))[1] and immature neurons continue to migrate tangentially from the subventricular zone (SVZ) to the olfactory peduncle and frontal cortex which continues for up to 18 postnatal months [Sanai et al., 2011].

Starting at 23 weeks, synapses emerge in the cortical plate [Molliver, Kostovic, Loos, 1973] and during the third trimester, neurons in the subplate form transient synapses with thalamic afferent neurons [Kostović, Judaš, 2010; Hevner, 2000] and short-range connections throughout the brain increase in number [Dehaene-Lambertz, Spelke, 2015].

[1] See fig. 2.3 for the time course of cortical layer development and notation.

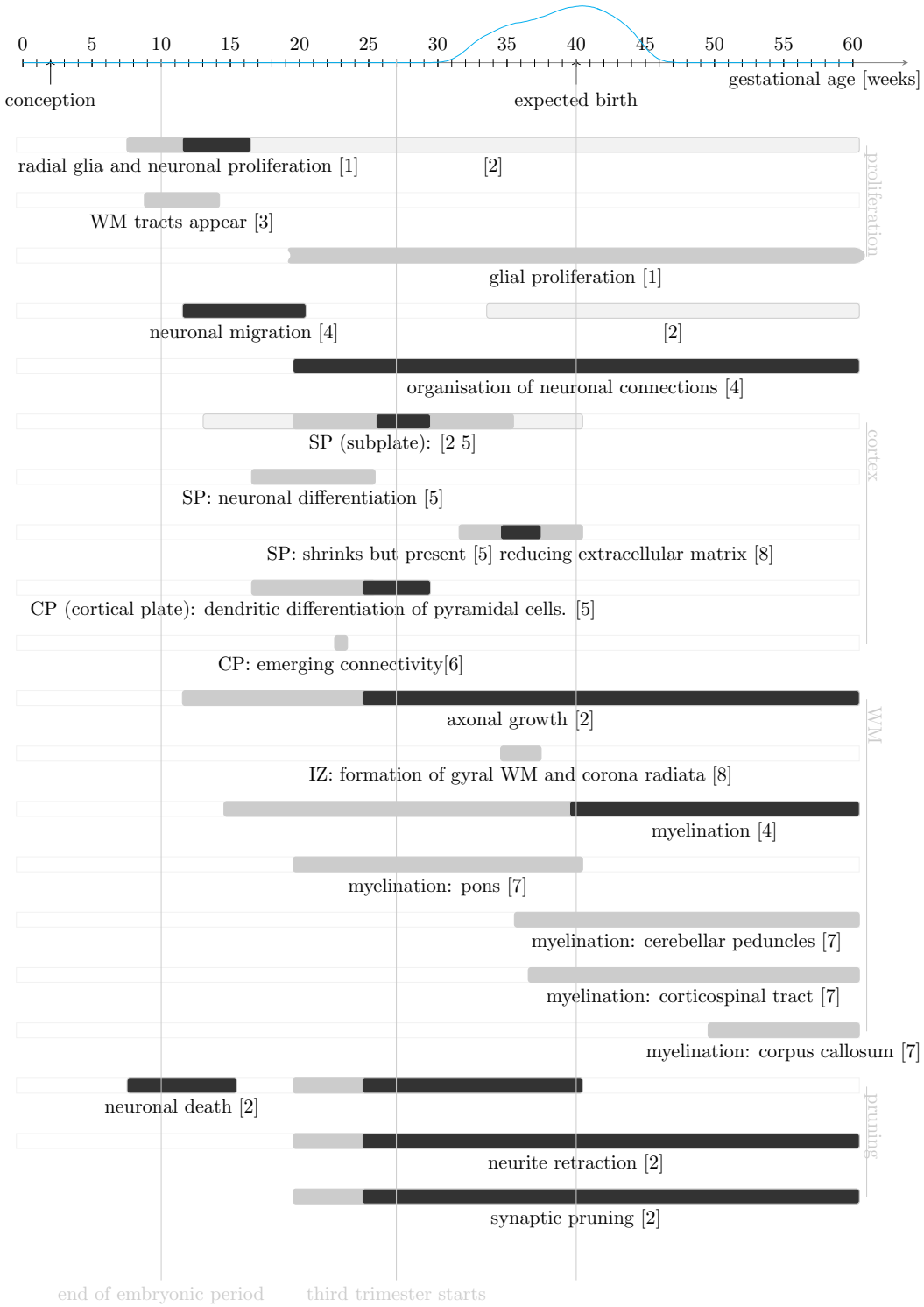


Figure 2.1.: Timeline of developmental events. The cyan line (top) shows the age distribution of the longitudinal cohort investigated in chapter 9. This diagram collects peak occurrences and timespans of developmental processes that I found in the literature. Darker shades mark peak occurrences. Note that although it incorporates diagrams from multiple reviews ([Andersen, 2003; Dubois et al., 2014; Knuesel et al., 2014; Linderkamp et al., 2009]), this diagram is not complete. For instance, Dubois et al. remark that "information on the beginning and ending of axonal pruning are missing in the human brain." References: [1]: [Volpe, 2008, chapter 2], [2]: [Linderkamp et al., 2009], [3]: [O'Rahilly, Müller, 2006], [4]: [Volpe, 2008, chapter 1], [5]: [Mrzljak et al., 1988], [6]: [Burkhalter, Bernardo, Charles, 1993], [7]: [Dubois et al., 2014], [8]: [Kostović, Jovanov-Milošević, 2006]

2.2.2. Neurons and neuroglia: basic units of brain parenchyma

2.2.2.1. Neurons

Neurons and glial cells are the basic cell type families that make up the majority of the brain parenchyma and determine its information processing function through complex electric and chemical processes [Debanne, 2004].

The number of named neuronal cell types goes into the hundreds [Masland, 2004] and neurons are heterogeneous in their morphology, physiology and molecular fingerprint, have highly complex lineages, and neuron classification hierarchies vary between locations in the brain and across species [Zeng, Sanes, 2017]. Neuron properties are collected in databases [Ascoli, Donohue, Halavi, 2007] and the process of classifying cells remains far from complete (see [Zeng, Sanes, 2017]).

Zeng, Sanes categorise cortical neurons based on their function in excitatory neurons and inhibitory interneurons. Interneurons transmit information between different types of neurons, in contrast to motor and sensory neurons. Although less valuable for classification [Zeng, Sanes, 2017], the morphological properties of neurons are likely the most relevant from a diffusion perspective.

Given that even the basic structure of axons is heterogeneous, I will briefly describe a common structure to introduce the basic notation used in diffusion microstructure modelling. Neurons consist of a cell body (soma), dendrites and an axon. Dendrites (déntro, Greek for tree) are tree-like protrusions that gather information (post-synaptic potentials) and transmit it to the soma. This information is processed and integrated at the protrusion of the neuron ("axon hillock", located at the soma or at a dendrite) that leads into the axon [Debanne, 2004]. Axons transport electric impulses to the local neural cell network typically over distances of μm to mm but can project into distant structures over cm reaching $1m$ in humans. Axons and dendrites are collectively referred to as neurites or neuronal processes.

In humans, each neuron is programmed to develop up to one axon [Parker et al., 2013] but axons can bifurcate into multiple axon collaterals giving rise to recurrent networks. Axons and axon collaterals branch in their terminal area into multiple telonderia and can, therefore, reach many target sites [Hall, 2015, chapter 45]. Axon collaterals can connect back to the soma or reach other neurons in the close vicinity. Recent research shows that axonal trees not only transmit impulses but also themselves contribute to the information processing by regulating the propagation of information or even reversing the direction in which the action potential travels [Dehaene-Lambertz, Spelke, 2015]. In the hippocampus, this axonal and dendritic arborisation gives rise to networks of local information feedback loops (recurrent networks) that play a role in pattern completion and separation, spatial context and episodic memory [Freund, Buzsáki, 1996; Rolls, 2013].

Although the neuroarchitecture of the hippocampus is comparatively simple [Stevens, 1998], the morphological categorisation of hippocampal interneurons gives rise to at least 16 morphological classes based on the location of the soma, the most prevalent orientation of the dendrites with respect to the cortical surface (horizontal, vertical or star-shaped: stellate) and the cortical zone reached by the axon [Parra, Gulyas, Miles, 1998]. See

fig. 2.2 for a sample of the morphological and scale diversity of neurons across species and neuronal cell types.

Pyramidal neurons (see fig. 2.2 D) are a class of excitatory neurons, which make up the majority of neurons in the human cortex but can also be found in subcortical structures, particularly the hippocampus and amygdala [Spruston, 2008]. They are named for their pyramidal shaped soma from which long reaching dendrites protrude off the pointy (apical) end while dendrites on the opposing basal end of the soma are comparatively short. Compared to inhibitory neurons, pyramidal neurons' axons are linear, less intricate and connect to fewer and more distant cells [Huang, Di Cristo, Ango, 2007]. Pyramidal neurons have in common that they possess physiologically distinct dendritic domains, yet, within the group of pyramidal neurons, function and appearance are varied [Spruston, 2008].

2.2.2.2. Neuroglia

In the adult brain, different types of glial cells exist: astrocytes, oligodendrocytes, microglia and ependymal cells. The function of glial cells involves giving structure to neurons by guiding them during development, supplying neurons with nutrients, electrical insulation, clearance of dead and diseased neurons and they play a crucial part in chemical signalling and metabolism [Jäkel, Dimou, 2017].

The star-shaped astrocytes are the most numerous cell type in the adult brain [Kettemann, Ransom, 2005] and their roles include metabolic interaction with neurons, the regulation of water and ion concentrations and in the formation of the blood-brain barrier to name a few [Kimelberg, 2010]. Ependymal glial cells separate the ventricular system from the brain parenchyma and regulate exchange between the two systems. They produce, absorb and move the cerebrospinal fluid (CSF). Microglia are immunocompetent cells and constantly and rapidly move their fine extensions (filopodia) to sense their environment[Jäkel, Dimou, 2017].

NG2-glial cells are oligodendrocyte precursor cells that build functional synapses with neurons that only allow communication from the neurons to the NG2-glial cells [Sun, Dietrich, 2013]. NG2-glial proliferate into mature oligodendrocytes but their population persist throughout life by rapid self-regeneration [Hughes et al., 2013], which makes them stand out in the adult parenchyma [Dimou et al., 2008].

Non-myelinating oligodendrocytes are present in large numbers in the cortex but their function is not well understood [Jäkel, Dimou, 2017]. Myelinating mature oligodendrocytes give rise to lipid-rich membranes extrusions that wrap around axons to electrically insulate and feed them [Nave, 2010]. For details on myelination see section 2.2.4.2.

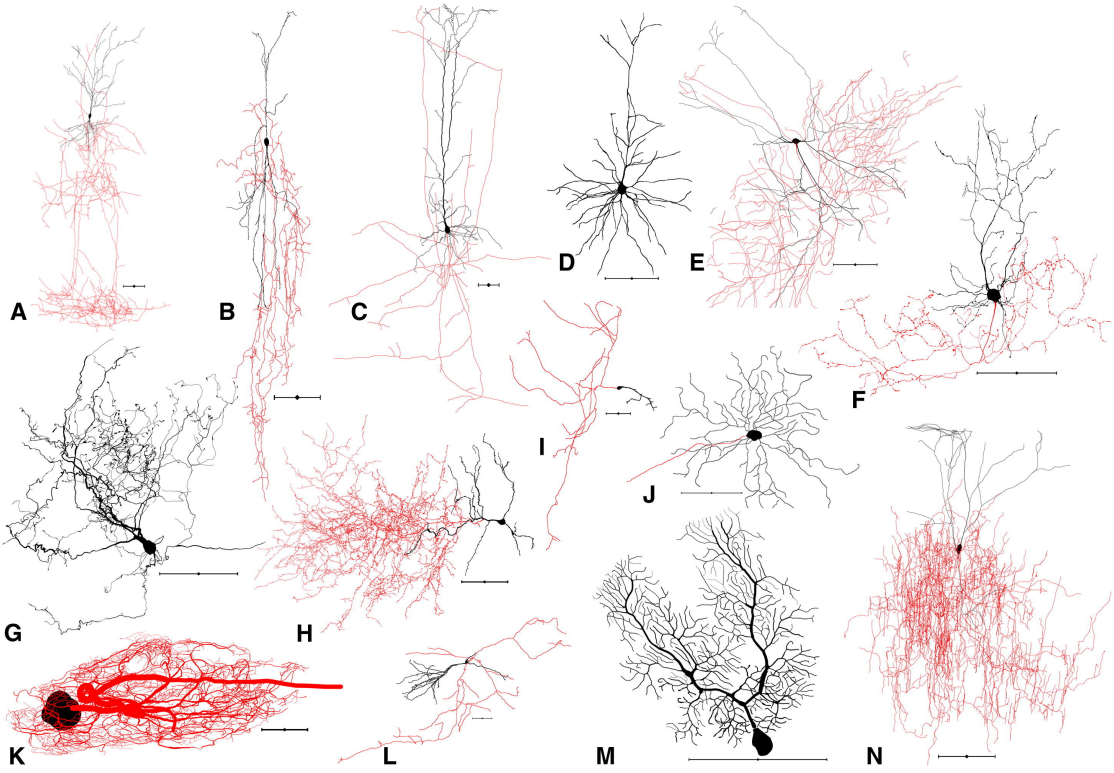


Figure 2.2.: "Morphological Diversity: A Representative Sample of Reconstructed Neurons from NeuroMorpho.Org (A) Rat neocortex Martinotti cell (NMO_00351). (B) Rat neocortex bipolar cell (NMO_06144). (C) Rat neocortex pyramidal cell (NMO_05729). (D) Mouse neocortex pyramidal cell (NMO_05549). (E) Mouse hippocampus Schaffer collateral-associated neuron (NMO_07893). (F) Mouse cerebellum Golgi cell (NMO_06902). (G) Cat brainstem vertical cell (NMO_06171). (H) Rat olfactory bulb deep short-axon cell (NMO_06222). (I) Mouse neocortex Cajal-Retzius cell (NMO_07521). (J) Mouse retina ganglion cell (NMO_06379). (K) Spiny lobster stomatogastric ganglion motoneuron (NMO_06635). (L) Rat hippocampus granule cell (NMO_06778). (M) Mouse cerebellum Purkinje cell (NMO_00865). (N) Rat neocortex layer 2/3 interneuron (NMO_04548). Scale bars represent 100 μm; somata and dendrites: black; axons: red." Reproduced with permission from [Parekh, Ascoli, 2013].

2.2.3. Neural proliferation and migration

The birth ("proliferation") and migration of neuronal cells in the brain starts at 7 weeks of gestation and the majority of neurons reach their final locations by the end of the second trimester (around week 20-24) [Rakic, 1990; Rakic, 2003]. Recent research has uncovered neurogenesis and migration, albeit shorter-range, lasting into adulthood [Ghashghaei, Lai, Anton, 2007]. It is hard to overstate the importance of this period for brain development as it defines the morphology and programs the future functional development of

the brain. Rakic has suggested that the roughly 1000-fold higher cortical surface area in humans compared to mice is caused by an additional 7 cell division cycles of neuronal progenitor cells in the human embryo [Rakic, 2009].

The early human brain consists of a single layer (neuroepithelium), which subsequently splits into separate zones, some of which disappear or transform into further separable layers later in gestation (see fig. 2.3). The zone that is in contact with the cerebrospinal fluid is called the ventricular zone (VZ), above which the transient "preplate" layer develops from week 10 [Honig, Herrmann, Shatz, 1996]. By week 14, the cerebrum can be divided into six zones starting from the ventricles: VZ, subventricular zone (SVZ), intermediate zone (intermediate zone (IZ)), subplate (SP), cortical plate (CP) and the marginal zone (marginal zone (MZ)), which is closest to the surface-facing membranes of the brain (the pial surface).

Early neuronal progenitor cells (specialised neuronal stem cells) are located in the ventricular zone (VZ, see fig. 2.3), and extend into the direction of the pial surface but might not reach it. They continue to divide into two identical progenitor cells and define the rate at which later proliferation occurs.

Later in gestation, neural progenitor cells in the ventricular zone of the developing brain divide into further specialised progenitor cells: neuroepithelial cells and radial glial progenitor cells. The latter are rooted in the VZ but extend to the pial surface. They subdivide and continue to divide into radial glial progenitor cells and neurons.

Two types of progenitor cells are located in the subventricular zone (SVZ), the cell layer adjacent to the VZ: the basal (a.k.a intermediate neuronal) progenitor cells, which do not reach the pial or ventrical surface; and recently discovered radial progenitor cells, which extend to the pial surface but do not reach into the ventricular zone [Lehtinen, Walsh, 2011]. The basal progenitor cells eventually give rise to most of the pyramidal projection neurons [Kowalczyk et al., 2009].

Neuronal proliferation occurs in two major phases. The first phase is characterised by radial glia and neuronal proliferation and occurs predominantly around 2 to 4 months of gestation [Clancy, Darlington, Finlay, 2001] and is followed by a phase of "glial multiplication", which lasts into the first year of life [Volpe, 2008, chapter 2].

Most neurons are created in the VZ [Volpe, 2008, chapter 2]. The SVZ gives rise to most glial cells [Volpe, 2008, chapter 2] and in later stages to neurons that terminate in the cortical layers [Kowalczyk et al., 2009]. The VZ disappears before birth and the intermittent zone (IZ) becomes the white matter [Honig, Herrmann, Shatz, 1996].

The majority of neuronal migration occurs between the third and fifth month [Volpe, 2008, chapter 2] but lasts into the perinatal period. Early migration is dominated by neurons extending their membrane radially from their place of birth to the cortex, which allows them to move their soma along these relatively short distances. When the cortex has grown further, two major directions of neuronal migration emerge: radially and tangentially with respect to the surface of the cortex. This migration occurs in distinct waves that are characterised by changes in neuronal shape or their direction of migration [Kriegstein, Noctor, 2004; Rakic, 1990]. Radial migration gives rise to projection fibres and the structure of the cortex.

Radial glial cells span a network of locally parallel fibres ranging from the VZ to the

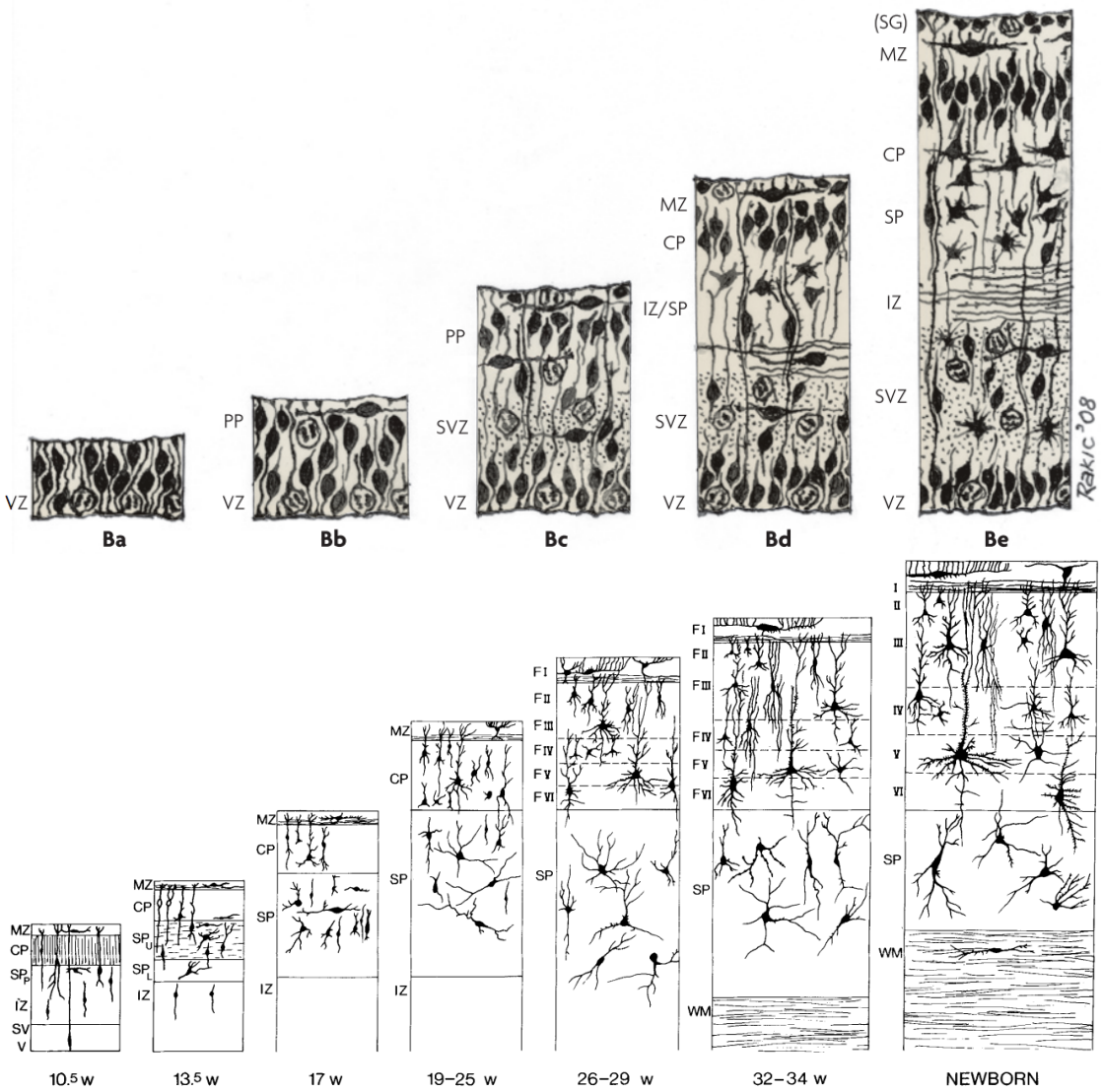


Figure 2.3.: Top: Revised version of the 1970s Boulder Committee's model of human neocortical development at 30 (a), 31-32 (b), 45 (c), and 55 (d) embryonic days and at 14 weeks (e). Adjusted from [Bystron, Blakemore, Rakic, 2008] with permission. Bottom: Schematic of cortical development between 10.5 weeks and birth. Reproduced from [Mrzljak et al., 1988] with permission. V(Z): ventricular zone, PP: preplate, SV(Z): subventricular zone, IZ: intermediate zone, SP: subplate zone, CP: cortical plate, MZ: marginal zone, which includes the subpial granular layer (SG), WM: white matter. For scale and locations of cortical zones see figs. 2.4 and 2.7.

cortical plate and serve as "scaffolding" for migrating neurons. In humans, they guide up to 30 generations of neurons simultaneously [Rakic, 1990] on their up to $7mm$ long radial migration to the cortex [Kriegstein, Noctor, 2004]. Neurons born early terminate

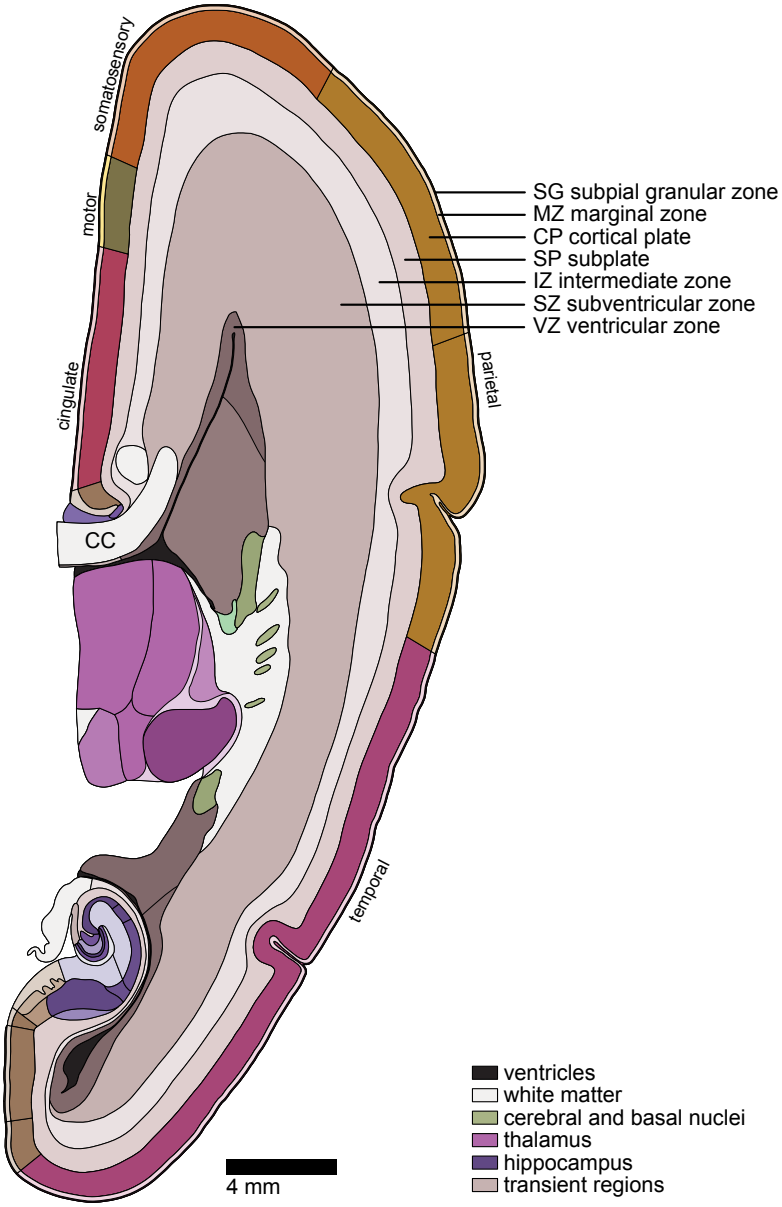


Figure 2.4.: Coronal schematic section of the human cerebrum at 21 weeks with annotated prenatal cortical layers. Adapted from the BrainSpan Atlas of the Developing Human Brain [Internet]. Funded by ARRA Awards 1RC2MH089921-01, 1RC2MH090047-01, and 1RC2MH089929-01. ©2011. Available from: http://brainspan.org

in the lower layers of the cortical plate and younger neurons migrate through the cortical plate extending it towards the MZ (see figure fig. 2.5, forming so-called neuronal columns that are believed to be crucial for information processing and abstraction [Tsunoda et al.,

2001].

At the same time, neuronal progenitor cells inside the SVZ, the IZ and the MZ migrate tangentially, guided by molecular mechanisms. The tangential migration of neuronal progenitor cells and cortical interneurons is followed by radial migration into the cortical plate [Volpe, 2008, chapter 2][Ghashghaei, Lai, Anton, 2007]. A recently identified proliferative zone in the ventral ganglionic eminences (which later form the basal ganglia) gives rise to interneurons that also migrate tangentially [Anderson et al., 2001].

Individual neurons can travel for weeks in the human brain [Rakic, 2003] and migration pathways vary between brain regions [Corbin, Nery, Fishell, 2001]. Furthermore, migration can occur in multiple stages and in different directions. For systematic reviews and details on neuronal migration see [Ghashghaei, Lai, Anton, 2007; Sidman, Rakic, 1973; Valiente, Marín, 2010; Kriegstein, Noctor, 2004; Marín, Rubenstein, 2001].

The subplate plays a crucial role in cortical development, particularly the neocortex in human and primates [Kanold, Luhmann, 2010]. Subplate neurons are heterogeneous in shape (see [Kanold, Luhmann, 2010]) and function and consist of at least 5 morphological subtypes. They are relatively mature in their dendrite density and synaptic innervation. Their dendrites can reach up to $1mm$ [Kanold, Luhmann, 2010].

The developed thalamocortical and corticothalamic pathways propagate sensory and motor information via the thalamus to cortical layer 4 and feed the signal from cortical layers 5 and 6 back into the thalamus (see fig. 2.6). These pathways are created in two stages, starting at the end of the second trimester and finishing around week 26 [Kostović, Jovanov-Milošević, 2006]. Thalamocortical axons first connect to neurons located in the transient subplate zone that are connected to the cortex (see fig. 2.5). This connection lasts 4 weeks after which thalamocortical axons extend into the cortex with locations determined by the previous interactions with subplate neurons. Corticothalamic migration is similarly divided by a transient connection via subplate neurons. In humans, subplate neurons undergo apoptosis after the establishment of the cortical connection. See section 2.2.4 for further discussion of the subplate.

By week 20 to 24 of gestation, most neuronal migration to the cerebral cortex has come to a halt [Rakic, 1990; Rakic, 2003] with only a few remaining areas of neurogenesis in the SVZ and the dentate gyrus of the hippocampus, and sparse and mostly short-range neuronal migration remaining into the second year of life and even fewer into adulthood [Cayre, Canoll, Goldman, 2009; Sanai et al., 2011; Ghashghaei, Lai, Anton, 2007]. However, neuronal maturation is far from complete. At birth, neurons in the frontal cortex are still small and poorly connected and have small and sparse dendritic trees (see fig. 2.6) [Courchesne et al., 2007]. Glial cells are thought to mediate neuronal maturation [Stiles, Jernigan, 2010] but the time course of neuronal maturation and glial differentiation and maturation in the neonatal and following postnatal period is still poorly understood, due to the limited post-mortem human brain samples from the postnatal period [Stiles, Jernigan, 2010; Dubois et al., 2014].

Glial progenitors keep proliferating and migrating beyond the time of birth, and in the case of oligodendrocyte progenitor cells continue throughout the lifespan. Glial progenitor cells reproduce in the SVZ of the forebrain and migrate mostly radially and differentiate in their target sites into astrocytes and oligodendrocytes [Cayre, Canoll, Goldman, 2009]. In

contrast to neurons, the majority of these processes happens peri- and postnatally [Stiles, Jernigan, 2010]. For a captivating review of oligodendrocyte precursor proliferation see [Richardson, Kessaris, Pringle, 2006].

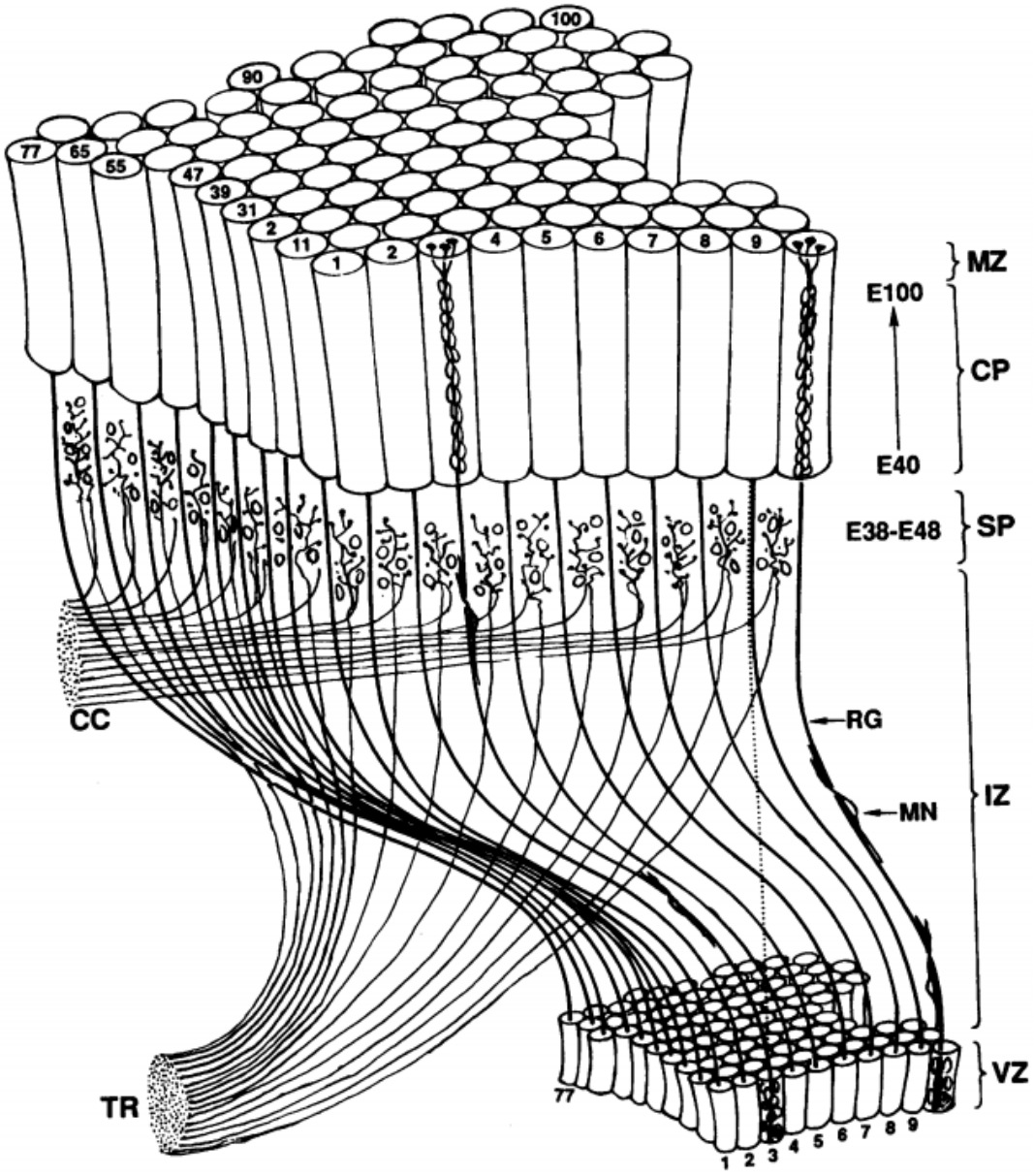


Figure 2.5.: "The relation between a small patch of the proliferative, ventricular zone (VZ) and its corresponding area within the cortical plate (CP) in the developing cerebrum. ... Neurons produced between E40 and EIOO by a given proliferative unit migrate in succession along the same radial glial guides (RG) and stack up in reverse order of arrival within the same ontogenetic column. Each migrating neuron (MN) first traverses the intermediate zone (IZ) and then the subplate (SP), which contains interstitial cells and "waiting" afferents from the thalamic radiation (TR) and ipsilateral and contralateral cortico-cortical connections (CC). ..." Reproduced from [Rakic, 1988] with permission.

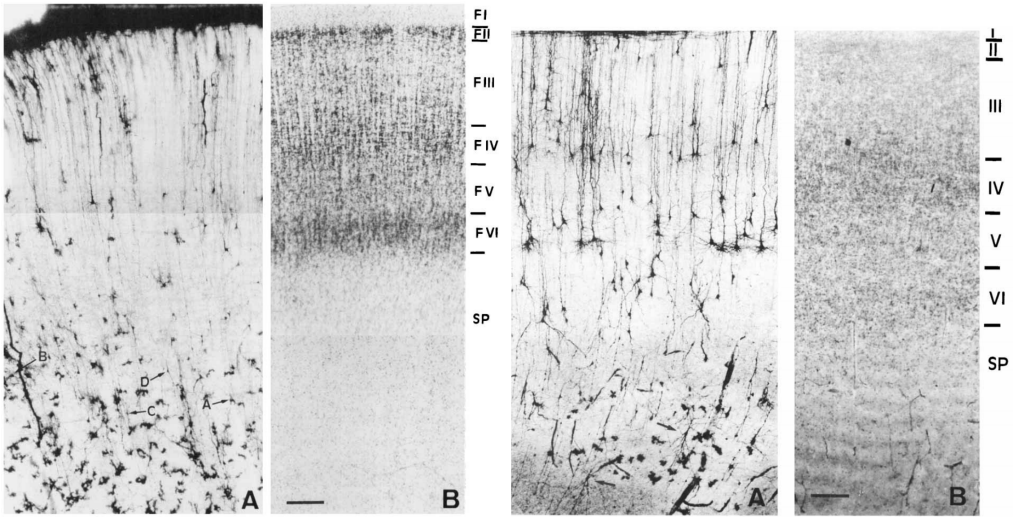


Figure 2.6.: Rapid Golgi stained (A) and Nissl (Cresyl Violet) stained (B) sections of the medial frontal gyrus of a 32 week old preterm born infant (left) and of an infant born at term (right). Golgi stains visualise the peripheral membrane of a sparse subset (1-3%) of neurons [Zaqout, Kaindl, 2016]. The relative sparsity of cells stained allows resolving the shape and extent of individual axonal and dendritic processes [Kobayashi, Lavenex, 2014]. Nissl staining colours Nissl substance in the majority of glial cells and neurons found in brain tissue. Nissl substance is found in nucleic acid (RNA and DNA) and extra-nuclear RNA (for instance in ribosomes). Hence cytoplasm and nuclei stand out in this stain, which is therefore commonly used to quantify cell body densities. Arrows in (A) point at the location of individual subplate neurons. Scale bars: 150 μm. Reproduced from [Mrzljak et al., 1988] with permission.

2.2.4. White matter development

White matter development involves the creation, migration, settlement and differentiation of neurons that subsequently form long-range connections in the brain and into the rest of the body. White matter is named after its appearance relative to grey matter; it is brighter due to the higher proportion of lipid-rich oligodendrocyte membranes that ensheath white matter axons ("myelin"). The majority of myelination occurs in the third trimester and postnatally and is ongoing into adulthood [Kinney et al., 1994; Brody et al., 1987].

Adult white matter contains besides neurons large numbers of glial cells (astrocytes, oligodendrocytes, microglia) and is categorised by the areas it connects. Projection fibres establish bi-directional connections between the cortex and the thalamus, the brainstem or the spinal cord. Commissural fibres establish inter-hemispheric connections, with the corpus callosum (corpus callosum (CC)) being the most prominent commissural white matter structure. Associative fibres form local connections between gyri ("u-fibres") or intra-hemispheric long-reaching cortical connections.

White matter development follows temporal trajectories that vary spatially, with structures responsible for sensory processing maturing before associative areas. Mapping those maturation patterns and their implications in humans is an active field of research of great clinical importance [Volpe, 2008] but with large gaps in knowledge [Dubois et al., 2014]. This is in part due to the unique developmental characteristic of human foetuses, and the challenges of fetal and neonatal imaging, and the lack of reliable and quantitative microstructural measures for in-vivo imaging [Dubois et al., 2014].

Early developing white matter tracts appear between 9 and 47 post-fertilisation days [O'Rahilly, Müller, 2006] and neurogenesis for most white matter neurons peaks before the tenth week [Bayer et al., 1993] and is mostly complete by week 16 [Clancy, Darlington, Finlay, 2001] (see section 2.2.3).

The majority of long-range connections are established in the second and third trimester. During the last trimester, white matter neurons develop dendritic connections with neurons in the adjacent grey matter and grow their axonal extension, giving rise to macroscopically aligned axon bundles. Axon growth and migration are guided by the routes defined by early matured axons, and neuronal activity, and proximity-dependent chemical factors [Dubois et al., 2014].

Neurons produced between 38 and 48 post-fertilisation days grow axons that form inter- and intra-hemispheric cortical connections into the subplate and connect the subplate with the thalamus (see section 2.2.3 and fig. 2.7). These early connections establish the organisation of white matter connections and guide later developing axons through their neuronal activity [Kanold, Luhmann, 2010] and their structural scaffolding [McConnell, Ghosh, Shatz, 1989]. Ipsilateral connections via axons from the developing corpus callosum reach the subplate between week 24 and 32 [DeAzevedo, Hedin-Pereira, Lent, 1997] and many long-ranging axons remain there until the subplate vanishes after birth [Kostović, Judaš, Sedmak, 2011]. In preterm-born babies, thalamocortical fibres are still growing into the subplate zone [Kostović, Jovanov-Milošević, 2006] and at term, short-range cortical connections continue to form in the subplate zone of the frontal

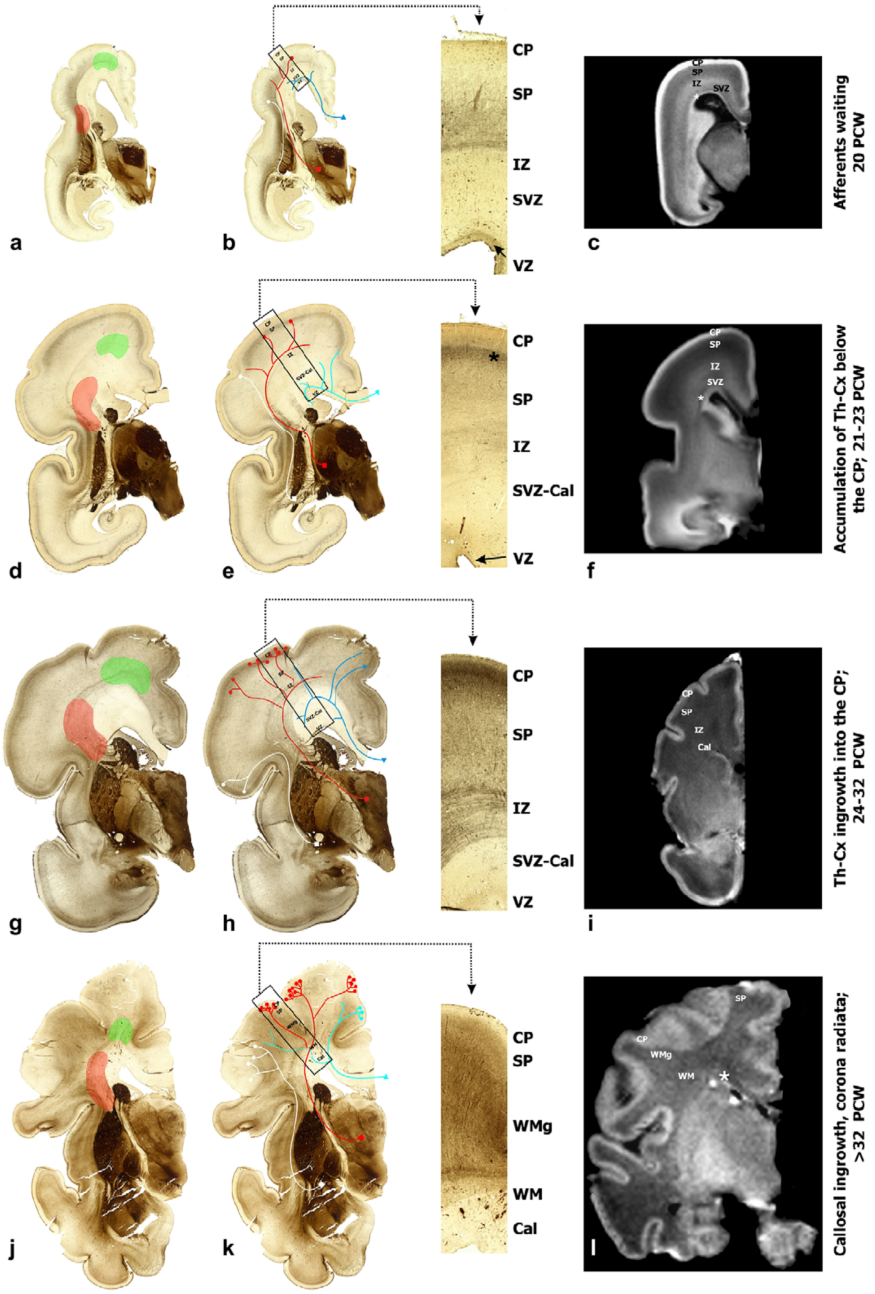


Figure 2.7.: "Transient fetal organization of the developing cerebral wall during the 20th post-conception week (PCW) (a, b, c), fetal phase (d, e, f), early preterm phase (g, h, i) and late preterm phase (j, k, l). The first main periventricular crossroads of pathways are shown in red, the second, frontal crossroads are shown in green (a, d, g, j). Growth of the cortical afferents (b, e, h, k): thalamocortical afferents are shown in red, callosal fibres in blue, and basal forebrain fibres in white lines. Insets illustrate the laminar organization of transient zones of the cerebral wall (from pia to ventricle). CP, cortical plate; SP, subplate zone; IZ, intermediate zone; SVZ, subventricular zone with cal, callosal fibres; WM, white matter; WMg, gyral white matter. Accumulation of the thalamocortical afferents in subplate is marked with a black asterisk (e). 'Retracting' callosal axons are shown as blue dotted lines (k). The transient zones are visible on in vitro MRI (c, f, i, l), as are fibre-rich periventricular zones (white asterisks in c, f, l)." Reproduced from [Kostović, Jovanov-Milošević, 2006] with permission.

cortex [Burkhalter, Bernardo, Charles, 1993].

In the superficial subplate areas called periventricular crossroads (see fig. 2.7), thalamocortical fibres and callosal fibres wait for triggers that determine the final migration to their connection point [Judaš et al., 2005]. The MRI properties of the fetal subplate are dominated by the extracellular matrix [Kostović et al., 2002; Judaš et al., 2005] and Kostović, Jovanov-Milošević predict that the cellular properties of the periventricular crossroads make it stand out in diffusion imaging relative to the other cortical layers [Kostović et al., 2002].

The extracellular matrix [Bosman, Stamenkovic, 2003; Novak, Kaye, 2000; Lau et al., 2013] refers to a part of the tissue that is found outside the cell membranes. It makes up about 20% of the adult brain volume [Nicholson, Syková, 1998], consists mostly of water [Cragg, 1979] but also contains large molecules, in particular saccharides (glycans) and proteins, through which it provides a dense mesh-like structural support for neuronal tissue in the brain ("perineuronal nets") [Kwok et al., 2011]. Note that the high water density and adhesive properties of the proteins make the extracellular matrix close to invisible in electron-microscopy images that do not counteract tissue shrinkage [Bignami, Hosley, Dahl, 1993]. It is involved in cell interactions and plays an important role in brain pathology [Lau et al., 2013; Horssen et al., 2006], repair [Sherman, Back, 2008] and development [Letourneau, Condic, Snow, 1994].

The viscosity of the extracellular matrix is presumed to be close to that of CSF [Nicholson, Syková, 1998] but the proximity to cell membranes causes a drop in apparent diffusion coefficient to about 38% the diffusivity of free water. The volume fraction of the extracellular matrix in the 10 day old rat cortex and corpus callosum is about 40%, twice as high as in adult rats [Bondareff, Pysh, 1968].

2.2.4.1. Regressive events

The rapid proliferation of neurons, their maturation and formation of synapses leads to an overproduction of neuronal connections, which is counteracted by subsequent programmed cell death (apoptosis) of neural progenitors and neurons and retractive and degenerative axonal and dendritic pruning [Riccomagno, Kolodkin, 2015]. These processes are subject to complex activity and molecular environment-dependent spatiotemporal interactions between neurons and between glial cells and neurons [Riccomagno, Kolodkin, 2015; Vanderhaeghen, Cheng, 2010].

All neuronal and neural progenitor cells undergo apoptosis and cell death is ubiquitous in the cortex during late gestation [Rakic, Zecevic, 2000]. The cortex has its highest number of neurons at 28 to 32 weeks of gestation, which drops by up to 70% by the time of birth [Rabinowicz et al., 1996]. Macroscopically, regressive events occur over timescales of weeks up to years; on the microscopic-level, development and pruning of individual cells happen in the timeframe of minutes up to hours [Stiles, Jernigan, 2010].

The newborn human and rhesus monkey brain does not form new axonal connections crossing the corpus callosum [LaMantia, Rakic, 1990; Kostović, Jovanov-Milošević, 2006] but axonal retraction is rapidly diminishing the number of axons in the corpus callosum [Innocenti, Price, 2005]. The corpus callosum of the developing rhesus monkey at birth

has more than 3.5 times the number of axons found in the adult brain and this number drops in the first three postnatal weeks by 43% at a rate of approximately 50 axons per second [LaMantia, Rakic, 1990]. This pruning succeeds the establishment of terminal zones of callosal fibres and is accompanied by an overall increase in synaptic connection in the cortex, which suggests that pruning is selective to excess ("supernumerary") neurons [LaMantia, Rakic, 1990]. However, the majority of neural apoptosis occurs before term [Stiles, Jernigan, 2010]

2.2.4.2. Myelination

In mammals, oligodendrocyte precursor cells are born in the ventral VZ in the spinal cord and in the forebrain from where they migrate into all parts of the brain and subsequently differentiate into oligodendrocytes [Richardson, Kessaris, Pringle, 2006; Baumann, Pham-Dinh, 2001]. The oligodendrocyte differentiation is followed by an increased production of myelin protein and the development of membrane protrusions that start wrapping loosely around axons. A single oligodendrocyte covers segments of multiple axons and a single axon is covered by segments of myelin sheaths interrupted by short myelin-bare segments called the "nodes of Ranvier" (see fig. 2.8). Over time, the number of wraps increases while the membrane structure compacts and chemically matures forming an increasingly dense and thick lipid-rich sheath. These processes combined are referred to as "myelination".

Adult myelin sheath The piece-wise electrical insulation the myelin sheath provides the axons with, allows neurons to change their mode of information propagation to saltatory conduction. In this form of conduction, action potentials do not need to be propagated via ion channels through the full length of the axon but they can "jump" over myelinated parts of the axon, which results in roughly 60-fold increased propagation velocity with low energetic and cross-sectional area demand. However, myelinated axons are not all heterogeneously covered by myelin [Tomassy et al., 2014]. Early proliferating oligodendrocytes tend to create long segments [Young et al., 2013] and segment length is spatially variable across the brain [Bakiri et al., 2011]. Variation in myelin thickness and node distance influences the conduction velocity [Waxman, Pappas, Bennett, 1972; Bakiri et al., 2011] and might, therefore, play a role in the fine control of information synchronisation [Fields, 2014]. Intermittent myelin-free axonal sections allow the formation of axonal synapses [Somogyi et al., 1998] and the release of neurotransmitters, which increases the degrees of freedom to the formation of signal circuits and processing within the neuron and between adjacent cells [Fields, 2014]. Not all axons myelinate but those that do tend to be thicker and, in turn, oligodendrocytes cause axons to grow in calibre [Friede, 1972; Sánchez et al., 1996; McTigue, Tripathi, 2008].

Mature myelin is composed of multiple dense layers with each layer having the cross-sectional molecular sequence: "protein-lipid-protein-lipid-protein" (see fig. 2.8) with 70% of the dry weight being lipids [Quarles, Macklin, Morell, 2006]. Electron microscopy shows protein layers as more electron dense, which makes adjacent protein-rich layers of low-magnification images appear as thick dark lines. Note that EM staining introduces

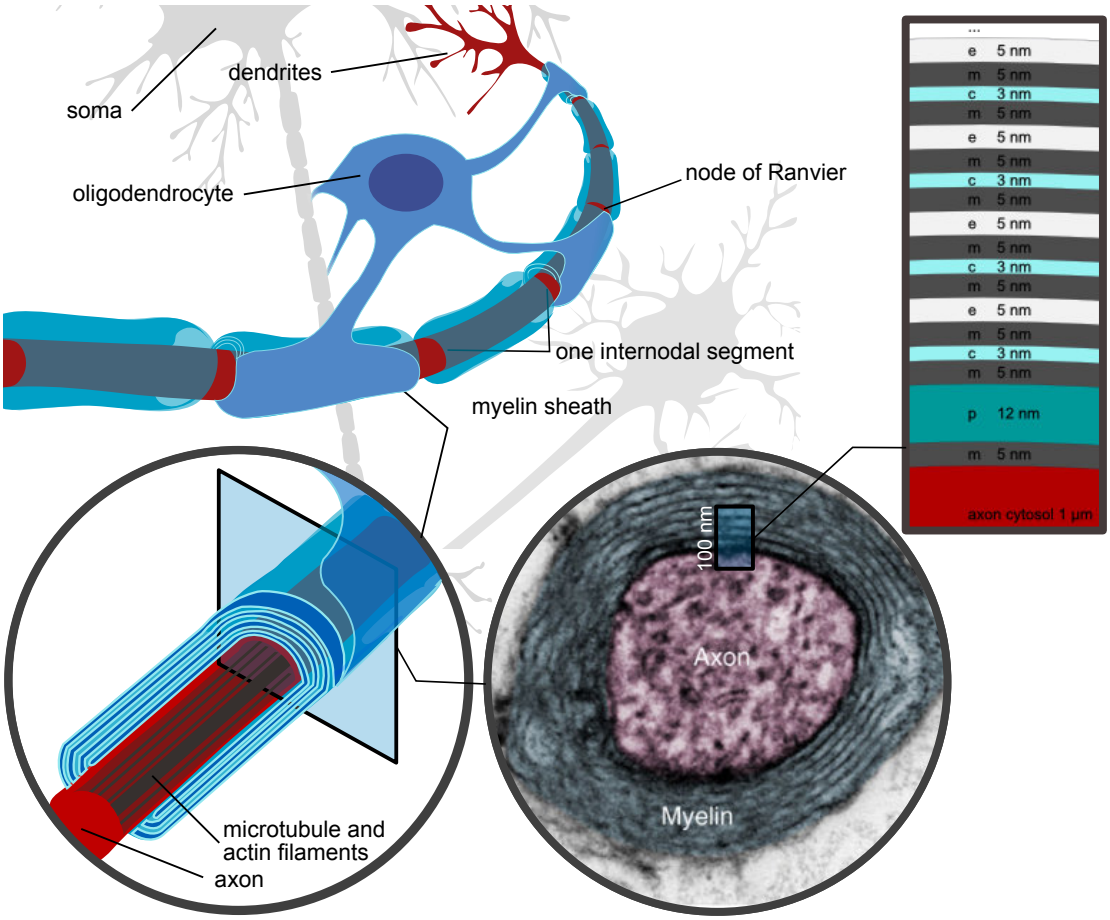


Figure 2.8.: Diagram of a myelinated neuron and its parts. The rectangular inset shows a portion of the axonal cytosol, as well as oligodendrocyte cytosol (c) and membrane (m), and extracellular (e) and pericellular (p) space drawn to scale. Myelin sheath thickness values and the adapted transmission electron image are from [Kwon et al., 2017] (Creative Commons CC BY), the neuron drawing is based on [Myelinated neuron, 2018] (public domain).

tissue compression and distortion that is not present in in-vivo X-ray or neutron scattering [Denninger et al., 2014] or optical scattering techniques [Schain, Hill, Grutzendler, 2014]. A myelin sheath has two cytoplasmic spaces, one that can be reached from the extracellular space without crossing a membrane and one that can be traced from the oligodendrocyte's cytosol (see fig. 2.8). The cytosol space has a depth of about $3nm$ in mice [Kwon et al., 2017] but its thickness is variable across species [Blaurock, 1981]. The extracellular space between myelin sheaths can swell by up to $8nm$ under osmotic pressure [Rand, Fuller, Lis, 1979].

If one were to unroll the oligodendrocyte protrusion that forms compact myelin, it would resemble a trapezoid-shaped pizza-like structure with a flat basin and bulges of cytoplasm at the inner- and distal edges parallel to the axon and at the sides adjacent to the nodes of Ranvier [Snaidero et al., 2014]. In contrast to peripheral myelin, the outermost wedge of cytoplasm parallel to the axon does not loop completely around the axon.

Oligodendrocyte development Similar to the neuronal overproduction, oligodendrocytes are produced in excess and undergo apoptosis if they do not receive signalling markers from adjacent axons [Raff et al., 1993]. However, the majority of oligodendrocytes are produced postnatally. Neurons control the survival of oligodendrocytes during development, which makes sure that the presence of oligodendrocytes is matched to the current requirement of myelinating axonal surface area [Barres, Raff, 1999]. Molecular and synaptic signalling [Lin, Bergles, 2004] and metabolic interaction and symbiosis between neurons and oligodendocytes continues into adulthood and is an active field of research [McTigue, Tripathi, 2008; Michalski, Kothary, 2015].

Oligodendrocyte development is a multi-staged process with distinct transient phases [Miller, 2002; Baumann, Pham-Dinh, 2001; Sherman, Brophy, 2005]. The cell lineage consists of oligodendrocyte precursor cells, immature oligodendrocytes, and finally, mature oligodendrocytes [Hardy, 1997] and these stages are of great importance for neuropathology related to prematurity [Back et al., 2001; Volpe, 2001]. In human white matter, oligodendrocyte precursor cells and immature oligodendrocytes can be found between week 18 and term but the latter cell population increases sharply from 10% to more than 30% of the oligodendrocyte population between week 27 and 30 [Back et al., 2001]. Back et al. defines oligodendrocytes as mature when they express "myelin-associated markers that include myelin basic protein (MBP)"[2]. This stage is first detected in parts of the periventricular white matter at week 30, where the density of mature oligodendrocytes markedly increases until birth [Back et al., 2001]. The genesis of immature oligodendrocytes before birth and their presence with a concurrent absence of myelin, which spans 3 months in humans, are not observed in rodents [Back et al., 2001; Reynolds, Hardy, 1997].

[2]This is in contradiction to other definitions [Watkins et al., 2008; Hardy, Friedrich Jr, 1996]. Watkins et al. find that, in cell-culture, immature oligodendrocytes start expressing myelin but oligodendrocytes loose this ability mostly when they mature [Watkins et al., 2008]

Myelin sheath development and maturation The development of the myelin sheath is predated with a pre-myelinating oligodendrocyte growing multiple thin processes for distances of typically $50\mu m$ that seek axons [Hardy, Friedrich Jr, 1996]. On contact with the axon, a process grows along the axon as a thin process and then develops membranes that envelop the axon once. This first part of the forming myelin sheaths grows along the axon while progressively forming more layers creating the swiss roll-shaped sheath of an internode (see [Snaidero et al., 2014] for a conclusive and complete study of this process). On contact with axons, the oligodendrocyte goes into a transitional stage during which it retracts the environment-sensing processes in the vicinity of successful "initiator processes" [Hardy, Friedrich Jr, 1996]. Myelin formation occurs primarily during a brief window early in oligodendrocyte development and is mediated by interactions between the oligodendrocyte and the axon and astrocytes [Watkins et al., 2008].

This early sheath has properties between that of the oligodendrocyte cell membrane and that of mature myelin [Barkovich, 2000] and undergoes biochemical changes during maturation [Kinney et al., 1994]. In general, mature membrane proteins are formed before lipids [Quarles, Macklin, Morell, 2006, chapter 4]. This increase of extracellular-space facing lipid content (mostly cholesterol and glycolipids) is a likely cause of T_1 relaxation times shortening and increased magnetisation transfer [Kucharczyk et al., 1994] observed in developing white matter [Barkovich et al., 1988; Barkovich, 2000].

The molecular synthesis occurs at the distal part of the process, which continues to grow around the axon and therefore increases the number of layers from the inside of the sheath [Sherman, Brophy, 2005]. While the sheath grows, it develops transient proteins (myelin-associated glycoprotein) that serve as structural support separating adjacent membranes, which gives rise to spaces in which proteins can form [Barkovich, 2000]. MBP is likely synthesised in the inner parts of the sheath and diffuses through the loosely wrapped sheath to the distal part of the sheath [Snaidero et al., 2014].

Concurrently, cytoplasm gets expelled and juxtaposed myelin layers gradually come into close contact through binding proteins (myelin proteolipid protein) located in the outer lipid bilayer [Coet, Suzuki, Popko, 1998]. This process starts in the outer layers and causes an inward-moving compaction of the myelin sheath and a reduction of free water content [Michalski, Kothary, 2015] (see [Snaidero et al., 2014] for high-resolution EM images). On a macroscopic level, maturing white matter contains decreasing amounts of unbound water, which is linked to a reduction of the free water T_1 and T_2 relaxation times and an increased compartmentalisation [Matsumae et al., 2001].

The final thickness of the myelin sheath scales with the thickness of the axon. The ratio of diameter between myelinated axon and the axonal membrane is called the "g-ratio" and ranges from 0.6 to 0.7 across animals [Sherman, Brophy, 2005]. However, neural activation during development and in adulthood can cause mature oligodendrocytes to reactivate the ensheathment process, which leads to neuroplasticity on the level of the myelin sheath [Gibson et al., 2014; Wake, Lee, Fields, 2011].

Spatiotemporal myelination patterns Also on the macroscopic level, myelination is a gradual progress during which myelin in a given white matter tract goes from absent

to only microscopically detectable on a myelin stained section, perceivable by eye, over to adult-like density (see fig. 2.9) [Gilles, Shankle, Dooling, 1983, chapter 12]. Gilles, Shankle, Dooling and Yakovlev, Lecours report this progression over time in great detail [Yakovlev, Lecours, 1967].

The maturation of myelin in white matter follows region specific sigmoidal patterns with variable onset and growth rates. In general, the progression of myelination in the human brain follows posterior-to-anterior, central-to-peripheral and inferior-to-superior trajectories [Dietrich et al., 1988; McArdle et al., 1987; Leipsic, 1901; Barkovich, 2000]. See [Gilles, Shankle, Dooling, 1983, table 12-7, figure 12-4] for a structure-resolved summary of onset, maturation transitions and peak changes of myelination between week 20 and 48. Maturation in many major white matter tracts levels off before two years of gestation. For reviews on this sequence during infancy, see [Brody et al., 1987; Kinney et al., 1988; Dean et al., 2015]. Kinney et al. clusters myelination trajectories into 8 groups based on the presence of myelin at term and the time when mature myelin is observed. This grouping is possibly linked to the time-course of cortical development [Guillery, 2005].

One of the first maturing structures are motor roots, which exhibit microscopic myelin as early as 20 weeks and mature rapidly to birth [Gilles, Shankle, Dooling, 1983; Yakovlev, Lecours, 1967]. About 50% of the structures investigated in [Gilles, Shankle, Dooling, 1983] reach mature myelin levels at week 40. For instance, the posterior limb of the internal capsule starts myelinating at week 32 and reaches mature density at birth. The anterior limb, however, starts myelinating later with microscopic myelin-forming at 38 weeks and it matures much slower. At term, the corticospinal tract (CST) myelination has progressed to mature levels in the midbrain and in the central parts of the corona radiata (see fig. 2.9). The corpus callosum starts myelinating at week 32 but progresses very slowly and remains in the microscopic stage at week 48.

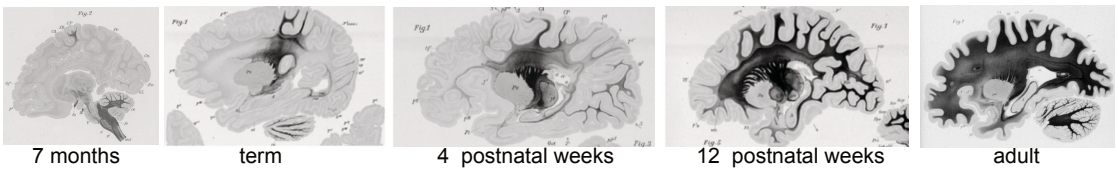


Figure 2.9.: Myelin and Nissl stained sagittal sections of the human foetus, term-born, infant and adult. Myelination follows a caudocranial, central to peripheral and posterior to anterior trajectory. The inferior cerebellar peduncles start myelinating before 25 weeks and myelination progresses in the adjacent brainstem and thalamic nuclei before it reaches primary motor and somatosensory, visual and auditory and finally association cortical areas [Fields, 2005; Barkovich et al., 1988]. Images are adapted and manually adjusted in contrast from [Dehaene-Lambertz, Spelke, 2015; Guillery, 2005; Flechsig, 2017], which have the common primary source [Flechsig, 1920] (public domain).

Chapter 3

Diffusion imaging of the brain

Contents

This chapter reviews the physical principles of diffusion, diffusion MRI, diffusion in brain tissue and methods to infer biophysical properties of brain tissue from the diffusion MRI signal.

3.1. Molecular diffusion, Brownian motion

In 1827, Robert Brown described the jiggling motion of microscopic pollen grains in fluid [Brown, 1828]. This random movement was first mathematically described by the statistician Thorvald Thiele [Thiele, 1880] and independently in 1905 by Albert Einstein [Einstein, 1905]. A series of experiments confirmed that water molecules move randomly, impacting the microscopic particles with random momentum and direction.

This random movement is referred to as diffusion and can be described using probability theory or macroscopic or microscopic physical principles [Vlahos et al., 2008; Philibert, 2005]. Here, I will briefly discuss diffusion from a macroscopic (Fick's laws) and microscopic (random walk) point of view.

3.1.1. Fick's law of diffusion

Diffusion in gas or liquids is a process that is driven by collisions of particles (such as water molecules) due to their thermal motion.

Adolf Fick introduced the laws of diffusion as an analogue phenomenon to the spread of heat in material caused by a heat gradient [Fick, 1855]. He adapted Fourier's theory of heat transfer by replacing local heat with local density or concentration $\rho(\mathbf{x}, t) = \rho$ and introduced a diffusion coefficient D similarly to a material's heat conductivity.

Fick's first law of diffusion states that the particle flux $\mathbf{J}$ (the rate at which particles traverse an infinitesimal area) is proportional to the concentration gradient:

$$\mathbf{J} = -D\nabla\rho(\mathbf{x}, t) \tag{3.1}$$

Using the conservation of mass $\frac{\partial\rho}{\partial t} + \nabla\mathbf{J} = 0$, one can derive Fick's second law, also called the diffusion equation, which states the rate of change in density over time

$$\frac{\partial\rho}{\partial t} = -\nabla\mathbf{J} = \nabla\left(D\nabla\rho\right) = D\nabla^2\rho \tag{3.2}$$

Assuming that all particles are located at $\mathbf{x} = \mathbf{0}$ in an unbounded d-dimensional space at time $t = 0$, the solution to 3.2 is

$$\rho(\mathbf{x}, t) = \frac{1}{(4\pi Dt)^{\frac{d}{2}}} e^{\frac{-\mathbf{x}^2}{4Dt}} \tag{3.3}$$

Einstein expressed Fick's laws as a probabilistic process in which the probability that a particle located at $\mathbf{x}_0$ at time 0 moves to location $\mathbf{x}$ at time t is described by the distribution $\mathcal{P}(\mathbf{x}, t|\mathbf{x}_0, 0)$.

Assuming an initial particle density $\rho(\mathbf{x}_0, 0)$, this yields the expectation value (the average over all possible values, weighted by their probability) of the local particle concentration at $\mathbf{x}$ and time t:

$$\rho(\mathbf{x}, t) = \int \rho(\mathbf{x}_0, 0)\mathcal{P}(\mathbf{x}, t|\mathbf{x}_0, 0)d\mathbf{x} \tag{3.4}$$

Using Fick's second law (eq. (3.2)), one can express the diffusion equation in terms of the conditional probability

$$\frac{\partial \mathcal{P}(\mathbf{x}, t|\mathbf{x}_0, 0)}{\partial t} = D\nabla^2_{|\mathbf{x}_0} \mathcal{P}(\mathbf{x}, t|\mathbf{x}_0, 0) \tag{3.5}$$

with $\nabla_{|\mathbf{x}_0}$, the gradient with respect to the initial condition.

If we assume all particles at time 0 to be located at x_0, $\mathcal{P}(\mathbf{x}, t|\mathbf{x}_0, 0)$ is equal to the Dirac delta distribution $\delta(\mathbf{x} - \mathbf{x}_0)$ and the solution to eq. (3.5) is

$$\mathcal{P}(\mathbf{x}, t|\mathbf{x}_0, 0) = \frac{1}{(4\pi Dt)^{d/2}} e^{\frac{(\mathbf{x}-\mathbf{x}_0)^2}{4Dt}} \tag{3.6}$$

3.1.2. Discrete-time random walk

This section introduces diffusion as a discrete-time step random-walk process, which is a Markov process. For more detailed and general derivations see [Bressloff, 2014, chapter 2].

Brownian motion can be interpreted as a random walk, which is a stochastic process that determines the direction and distance each particle – or water molecule – moves in a given discrete, constant time interval δt. Assuming a particle located at $\mathbf{x}_0 = \mathbf{0}$ at time $t_0 = 0$, the location of that particle after n time steps is the sum of steps it took:

$$\mathbf{x}_n = \sum_{i=1}^{n} \mathbf{s}_i. \tag{3.7}$$

Since we are interested in diffusion in an isotropic, viscous fluid in equilibrium, we'll assume that any step $\mathbf{s}_i$ is independent of the previous step $\mathbf{s}_{i-1}$ and that steps are random variables with values drawn from a common probability distribution $p(\mathbf{s})$. Hence, $\mathbf{s}_i$ are random variables.

A measure of the spreading of a particle is its mean-square-displacement (MSD). It is defined as the expectation value of the squared distance from the original position at time step n and can be calculated by squaring equation 3.7.

$$\begin{aligned} \mathrm{MSD}(\mathbf{x}_n) &:= \left\langle (\mathbf{x}_n - \mathbf{x}_0)^2 \right\rangle && (3.8)\\ &= \left\langle \sum_{i=1,j=i}^{n} \mathbf{s}_i^2 + \sum_{i,j=1,i\neq j}^{n} \mathbf{s}_i^T \mathbf{s}_j \right\rangle && (3.9)\\ &= \sum_{i,j=1,j=i}^{n} \left\langle \mathbf{s}_i^2 \right\rangle + \sum_{i,j=1,i\neq j}^{n} \left\langle \mathbf{s}_i^T \mathbf{s}_j \right\rangle && (3.10) \end{aligned}$$

Provided that the step distribution has zero mean (particles do not drift or flow), $< \mathbf{s}_i^2 >$ is the variance of $\mathcal{P}(\mathbf{s})$. $< \mathbf{s}_i^T \mathbf{s}_j >$ is the covariance between steps i and j, which has to be zero as steps are assumed to be independent.

Therefore in one-dimensional space, the MSD after n time steps with steps drawn from a distribution $\mathcal{P}(\mathbf{s})$ with variance $\langle \mathbf{s}_i^2 \rangle = l^2$ becomes

$$\mathrm{MSD}_{1D}(\mathbf{x}_n) = \sum_{i,j=1,j=i}^{n} \langle \mathbf{s}_i^2 \rangle = n \ \mathrm{var}(\mathcal{P}(\mathbf{s})) = nl^2 \tag{3.11}$$

and analogous, for 3-dimensional random walk, the MSD becomes

$$\mathrm{MSD}_{3D}(\mathbf{x}_n) = \sum_{i,j=1,j=i}^{n} \langle \mathbf{s}_i^2 \rangle = \sum_{i,j=1,j=i}^{n} \left\langle \sum_{d=1}^{3} (\mathbf{s_i}^T \mathbf{e}_d)^2 \right\rangle \tag{3.12}$$

$$= n \left(\sum_{d=1}^{3} \mathrm{var}(\mathcal{P}(\mathbf{s}^T \mathbf{e}_d)) \right) = n \left(\sum_{d=1}^{3} l_d^2 \right). \tag{3.13}$$

with $\mathbf{e}_d$ the unit vector of dimension d.

Using $t = n\tau$ with τ, the time between two collisions, the one-dimensional MSD is

$$\mathrm{MSD}_1(\mathbf{x}) = nl^2 = \frac{t}{\tau} l^2 = 2Dt, \text{ with } D := \frac{l^2}{2\tau}. \tag{3.14}$$

Note that the MSD is linear with time; equivalently, the root-mean-square-distance increases with the square root of the time.

For 3-dimensional random walk with equal probability distribution variance in any dimension($l_1 = l_2 = l_3 = l$) the MSD becomes

$$\mathrm{MSD}_{3D}(\mathbf{x}) = n3l^2 = 6Dt \tag{3.15}$$

For free water at 20°C and 37.5°C, D is $2.0\mu m^2/ms$ and $3.1\mu m^2/ms$, respectively[1].

3.2. Principles of diffusion weighted imaging

Before nuclear magnetic imaging was invented, Bloch and Purcell [Bloch, 1946; Purcell, Torrey, Pound, 1946] developed a spectroscopic technique called nuclear magnetic resonance (NMR), which applies to all atoms with non-zero nuclear magnetic moment ("spin") such as protons (^1H) found in water molecules.

The main idea behind ^1H-NMR is that by applying a strong stationary (homogeneous) magnetic field ($\mathbf{B_0}$, parallel to the z-axis), a state is created in which protons align in one of two configurations – parallel or antiparallel to the magnetic field – precessing at a field-dependent frequency and with a fixed "angle" relative to the $\mathbf{B}_0$-axis and with the majority of protons aligned with the magnetic field causing a net-magnetisation $\mathbf{M}_0$ in that direction. The NMR experiment measures the response of the spins to electromagnetic perturbations. This response depends on the density of the spins, the energy of the pulse, and the interactions of the spins with each other and their environment. This allows probing properties of material and tissue. An excitation pulse with a magnetic

[1] http://dtrx.de/od/diff/

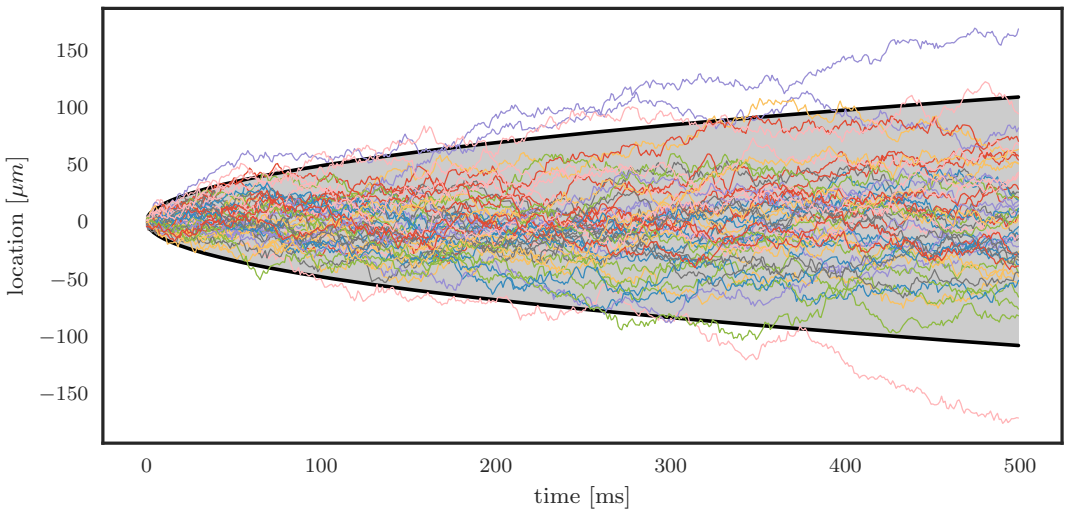


Figure 3.1.: Simulation of random walk of water molecules at 37.5°C showing the trajectories of 50 particles located at the origin at time 0 diffusing along the spatial axis over time. The grey area shows where 95% of the particles are expected to be found according to eq. (3.6).

field oscillating at the spins' precession frequency causes a subset of the spins to flip their orientation and to precess in phase with the excitation pulse. Their net magnetic moment can be measured as an NMR signal of a strength that depends on the proton density in the tissue. Over time the signal decays with tissue specific rates due to spin-lattice interactions (T_1-decay) and spin-spin interactions (T_2-decay, see section 3.2.2). NMR sequences can be designed to modulate these contrasts, giving it the ability to distinguish tissue types and normal from abnormal tissue.

Lauterbur [Lauterbur, 1973] introduced methods that use spatial encoding of the NMR signal, which allow the reconstruction of NMR images (MRI). Briefly, spatial information is inferred from the phase and frequency of the measured MRI signal. This encoding is typically imparted via gradient coils that cause known and distinctive, spatially-dependent, orthogonal, approximately linearly varying magnetic fields of specific duration and direction [Edelstein et al., 1980]. These spatially varying magnetic gradient fields (short "gradients") are superimposed onto the static $\mathbf{B}_0$-field when the spins are excited and before and while the signal is read. In 2D MRI, a linear gradient field is applied orthogonal to the slice plane, rendering the precession frequency of the spins a function of the (signed) distance to the slice. Using an excitation pulse with a limited bandwidth, it is possible to excite only those spins in the planes with corresponding precession frequencies. The slice thickness is a function of the bandwidth of the RF pulse and the gradient strength. For now, let us assume that the frequency spectrum is continuous within a defined bandwidth so that the excited planes make up a single contiguous slab (see section 3.2.3 for imaging multiple spatially separated slabs).

All spins in the excited slab contribute to the signal measured in the receiver coils as a sum of oscillating waves with different frequencies and phases. By applying a gradient during the acquisition of the signal, oriented perpendicular to the slice-selection gradient, the measured signal's frequencies can encode the location relative to this direction ("frequency encoding"). Similarly, a gradient applied in orthogonal direction to both gradients allows encoding the remaining spatial direction via a location-dependent phase shift ("phase encoding"). By repeatedly sampling the signal with varying phase-encoding, it is possible to transform the measured signals into a 2D image of signal densities via the Fourier transform.

Diffusion weighted imaging uses gradients to additionally encode statistical properties of the molecular movement that occurs in the sample between excitation and sampling with the spatial encoding of the NMR signal to form diffusion weighted images. For an overview of the physical foundations and the history of the field of diffusion MRI see [Price, 1997; Minati, Węglarz, 2007; Le Bihan, Johansen-Berg, 2012].

3.2.1. Pulsed gradient spin-echo (PGSE)

Hahn [Hahn, 1950] observed that the NMR signal amplitude of a spin-echo experiment (see fig. 3.2, top) is reduced if the spins are located in an inhomogeneous magnetic field. Following the initial excitation pulse, the inhomogeneous magnetic environment makes the spins precess at different rates, which causes a dephasing of the initially aligned spins. Disregarding spin relaxation processes, the 180° pulse subsequently applied after time TE/2, inverts this relative phase difference and causes the spins to realign at time TE when the NMR signal is measured. Yet, this is only true for stationary spins. Spins undergoing Brownian motion accumulate non-zero but random phase differences, which causes a reduction in the transverse magnetisation amplitude.

The Stejskal-Tanner sequence [Stejskal, Tanner, 1965], shown in figure 3.2, is a spin-echo sequence that uses two magnetic field gradients of amplitude G applied in the same direction $\mathbf{e}_G$ ($\mathbf{G} = G\ \mathbf{e}_G$) just before and after the 180° pulse.

Assuming that the gradient duration δ is short compared to the spacing between the leading edges of the gradients (Δ), we can treat spins as stationary during a gradient pulse and diffusing between the gradients ("narrow pulse approximation"). The first gradient causes a spatially dependent phase shift proportional to the gyromagnetic ratio γ (γ =2.675 10^8 rad s^{-1} T^{-1} for protons)

$$\phi_1(\mathbf{x}) = \gamma \int_0^t \mathbf{G}(t') \cdot \mathbf{x}_1 dt' = \gamma\delta\mathbf{G} \cdot \mathbf{x}_1 \tag{3.16}$$

after which the spin moves to location $\mathbf{x}_2$. The rephasing gradient induces the phase shift

$$\phi_2(\mathbf{x}) = \gamma \int_\Delta^{\Delta+\delta} \mathbf{G}(t') \cdot \mathbf{x}_2 dt' = \gamma\delta\mathbf{G} \cdot \mathbf{x}_2 \tag{3.17}$$

which results in a net phase shift of

$$\phi = \phi_1 - \phi_2 = \gamma\delta G\,(\mathbf{x}_1 - \mathbf{x}_2) \cdot \mathbf{e}_G \tag{3.18}$$

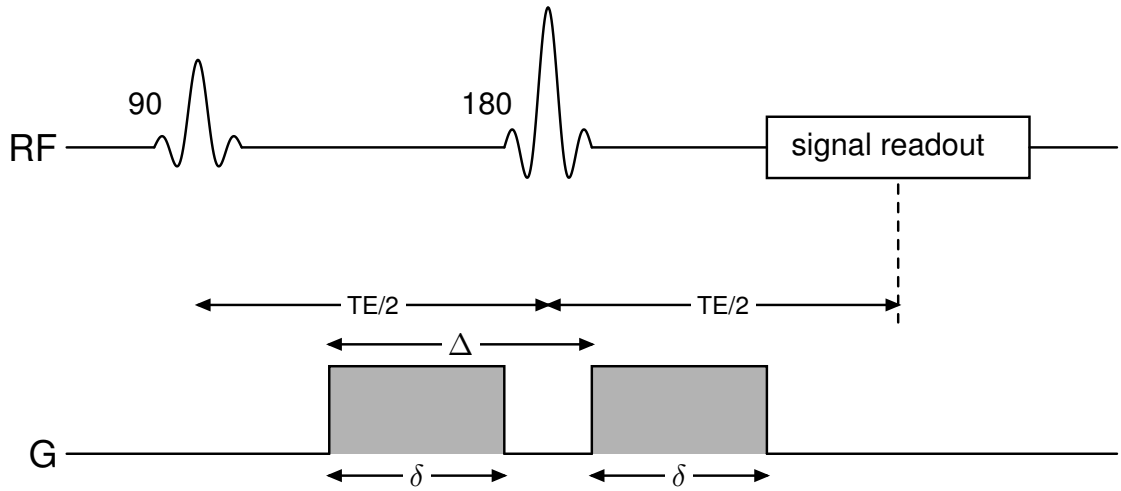


Figure 3.2.: Stejskal-Tanner spin echo sequence: The top line ("RF") shows a spin echo experiment where spins are excited with a 90° pulse. After dephasing for time TE/2, a 180° pulse inverts the phase relations for stationary particles which causes a zero phase-shift at signal readout time (TE relative to the first excitation).

If the spin did not change its position along the direction of the diffusion gradient, the phase difference (eq. (3.18)) becomes zero and the spin contributes fully to the echo signal. However, the phase shift linearly depends on the change in location along the gradient direction in the interval Δ and the product of diffusion gradient strength and duration. If spins undergo thermal motion, the ensemble average signal is attenuated due to dephasing:

$$S = S_0 \left\langle e^{i\phi} \right\rangle \tag{3.19}$$

with S_0 the signal without diffusion weighting gradients, where the average is taken over all initial spin positions and trajectories.

Using the probability distribution $\mathcal{P}(\mathbf{x}|\mathbf{x}_1, \Delta)$ for spins located at $\mathbf{x}_1$ moving to the location $\mathbf{x}$ in the time period Δ, we can express the diffusion attenuation in the narrow pulse approximation ($\delta \ll \Delta$) as

$$S = S_0 \int \int_V \rho(\mathbf{x}_1)\mathcal{P}(\mathbf{x}|\mathbf{x}_1, \Delta)e^{-i\gamma\delta\mathbf{G}\cdot(\mathbf{x}-\mathbf{x}_1)}d\mathbf{x}d\mathbf{x}_1 \tag{3.20}$$

with $\rho(\mathbf{x}_1)$ the initial spin density distribution in the volume of the object V, which is commonly assumed to be constant $(1/V)$ [Yablonskiy, Sukstanskii, 2010]. Hence, substituting $\frac{\gamma\delta\mathbf{G}}{2\pi}$ with $\mathbf{q}$, this becomes a 3D Fourier transform of the conditional probability function of displacements $\mathbf{s} = \mathbf{x} - \mathbf{x}_1$ with $\mathbf{q}$, the conjugate variable to $\mathbf{s}$ [Basser, 2002].

$$S = S_0 \int \mathcal{P}(\mathbf{s}, \Delta)e^{-2\pi i\mathbf{q}\cdot\mathbf{s}}d\mathbf{s} \tag{3.21}$$

Let the 1D coordinate system x' be defined along the diffusion gradient direction. Then for free Brownian motion ($\mathcal{P}(\mathbf{s},t) = \frac{1}{(4\pi Dt)^{1/2}} e^{\frac{s^2}{4Dt}}$, eq. (3.6)), eq. (3.21) becomes [Stejskal, Tanner, 1965]

$$S = S_0 \int \frac{1}{\sqrt{4\pi D\Delta}} e^{\frac{s^2}{4D\Delta}} e^{-i\gamma\delta G s} ds = S_0 e^{-\gamma^2\delta^2 G^2 \Delta D} \tag{3.22}$$

The diffusion weighting effect of the gradients and RF pulses in an NMR sequence on an object are commonly summarised in a single variable b, the "b-value" [Le Bihan, 1995]. For the Stejskal Tanner sequence with instantaneously switching gradients and $\delta \ll \Delta$ (fig. 3.2), the b-value is $\gamma^2\delta^2 G^2 \Delta$ (see eq. (3.22)), and the free diffusion signal decay becomes

$$S(b) = S_0 e^{-bD} \tag{3.23}$$

In fact, eq. (3.23) is valid for free diffusion with arbitrary gradient weighting causing a phase shift at time t

$$\phi(t) = \gamma \int_0^t dt' \mathbf{G}(t')\mathbf{r}(t') \tag{3.24}$$

if b is defined as [Karlicek Jr, Lowe, 1980; Le Bihan et al., 1986]

$$b = \gamma^2 \int_0^{TE} \left(\int_0^t \mathbf{G}(t') dt' \right)^2 dt \tag{3.25}$$

Taking a more realistic gradient trajectory with trapezoidal gradients with finite gradient duration and finite (but constant) ramp up and down slope into account, b becomes [Stejskal, Tanner, 1965]

$$b = \gamma^2\delta^2 \left(\Delta - \frac{1}{3}\delta \right) G^2 \tag{3.26}$$

See [Sinnaeve, 2012] for derivations of b-value for common gradient waveforms in Stejskal-Tanner sequences.

3.2.2. T_1, T_2 and T_2^* weighting

The Bloch equations [Bloch, 1946] describe how the sample magnetisation ($\mathbf{M}(t)$) changes over time in the presence of relaxation. Torrey formalised the effect of diffusion on the sample magnetisation by adding a diffusion term to the Bloch equations [Torrey, 1956]

$$\frac{d\mathbf{M}}{dt} = \gamma \left(\mathbf{M} \times \mathbf{B}_0 \right) + \left(\frac{\mathbf{M_x}}{T_2}, \frac{\mathbf{M_y}}{T_2}, \frac{\mathbf{M_0} - \mathbf{M_z}}{T_1} \right)^T + \nabla D \nabla \mathbf{M} \tag{3.27}$$

where T_1 and T_2 are the longitudinal and transverse relaxation times, respectively. The longitudinal relaxation refers to the return of excited spins to their original state by interaction with their surrounding environment ("lattice"). T_2 decay describes the loss of phase coherence due to spin-spin interactions in the sample.

The solution to eq. (3.27) after a 90° pulse is

$$M_{xy}(t) = M_0 e^{-\frac{t}{T_2} - bD} \tag{3.28}$$

This shows that the EPI signal is also modulated by an exponential decay with rate $1/T_2$. In practical experiments, the transverse signal decays at a faster rate than $1/T_2$ due to magnetic field inhomogeneities caused for instance by an imperfect $\mathbf{B}_0$-field or by heterogeneous magnetic susceptibility of the scanned object. The observed decay of the transverse signal is denoted by T_2^*, with $\frac{1}{T_2^*} = \frac{1}{T_2} + \frac{1}{T_2'}$ and T_2', the decay rate due to magnetic field inhomogeneities.

In addition, magnetic field inhomogeneities also contribute to the T_2-decay due to the random motion of diffusing particles through an effectively randomly varying magnetic field. This effect is present irrespective of any externally applied diffusion gradients and cannot be refocussed by a spin-echo.

Since diffusion-weighted images are typically acquired using spin-echo imaging, diffusion-weighted images are also inherently T_2-weighted.

3.2.3. Multi-slice spin-echo echo-planar imaging

Diffusion-weighted imaging requires large gradient amplitudes to achieve the required sensitisation to small random motion. This sensitisation to motion is a major problem for in-vivo imaging; babies in particular. Incoherent bulk motion or flow introduce phase errors across the object. Due to uncontrollable and irregular motion (particularly in pulsation-related motion), these errors result in inconsistencies in the data acquired between the different excitations.

Mansfield's echo-planar imaging (EPI) technique [Mansfield, 1977] allows acquiring all (single-shot EPI) or parts (multi-shot or segmented EPI) of the k-space required to form a 2D image after a single excitation by encoding the spatial information with rapidly alternating gradient pulses while the gradient-echo forms. The full readout typically takes on the order of 100ms. Single-shot EPI is particularly suitable for diffusion-weighted imaging as all data are acquired after a single preparation and therefore, all k-space samples are affected by the same phase errors, effectively freezing motion.

Figure 3.3 shows a schematic pulse diagram of the Stejskal-Tanner imaging sequence, in which the k-space is sampled during the readout using fast alternating gradient pulses applied in the frequency encode (x) direction and intermittent blips in the phase encode (y) direction, which shift to the next k-space line. This results in a zig-zag traversal of k-space.

To image the whole brain, multiple 2D EPI images can be acquired at different slice locations and stacked into a 3D volume. This is usually performed in an interleaved fashion by acquiring even and odd slices separately to minimise cross-talk or spin-history effects due to overlapping excitation pulses and motion between slices. The time for acquiring a single slice is determined by the duration of the diffusion preparation, the EPI-readout and additional delays due to thermal heating [Hutter et al., 2017]. The time between repeated excitations of the same slice (repetition time, TR) is linearly

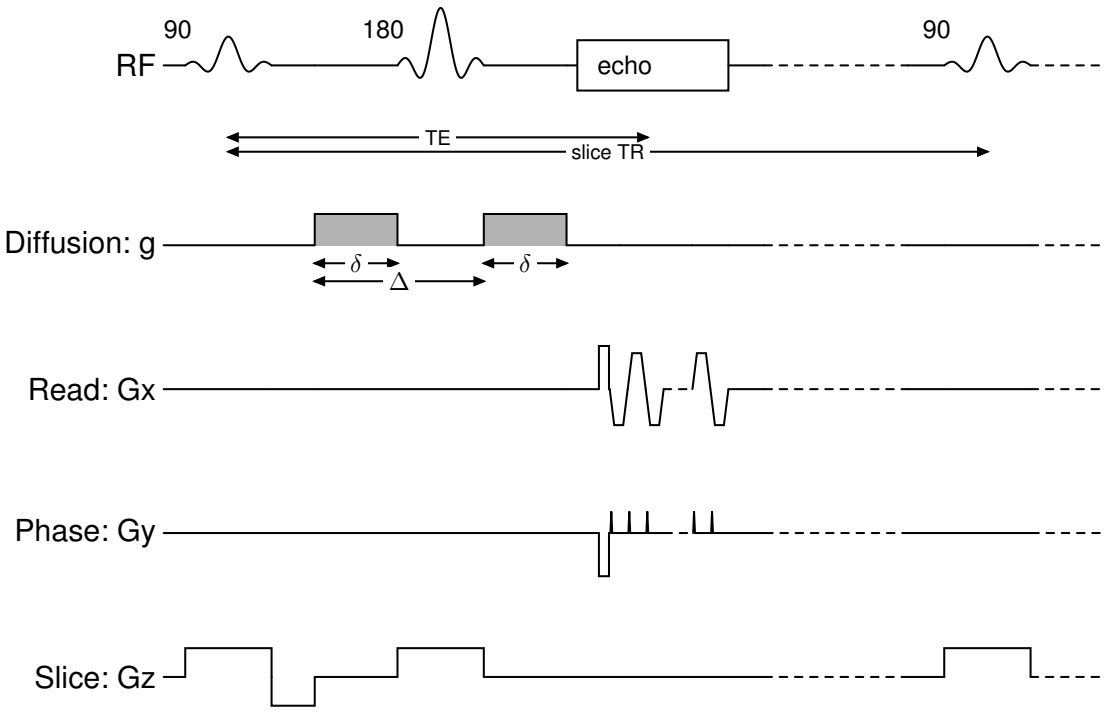


Figure 3.3.: Stejskal-Tanner spin echo imaging sequence with gradient duration δ and Δ, the delay between the onset of the two pulsed field gradients.

dependent on the number of slices. For volumes consisting of 50 to 60 slices, it takes on the order of 10s to acquire all slices. Inconsistent bulk motion during this time causes misalignment of slices and non-sampled areas, which are challenging or impossible to fix in post-processing.

To reduce the chance of motion corruption it is desirable to shorten the acquisition time. Partial Fourier techniques [McGibney et al., 1993] reconstruct the images from as little as half the k-space data, exploiting the conjugate symmetry of k-space. This technique is susceptible to phase errors, introduced for instance by motion or eddy currents, and therefore requires sampling significantly more than 50% of k-space for EPI sequences.

Parallel imaging methods allow a reduction of the time required to acquire a slice by reducing the number of phase encoding gradients. Using multiple imaging coils and information about their varying spatial sensitivities (SENSE [Pruessmann et al., 1999], GRAPPA [Griswold et al., 2002]), the signal can be reconstructed from the undersampled k-space data. However, the reduced amount of acquired data causes a reduction in SNR. Reduced SNR and imperfections in the signal separation due to insufficient differences and imperfections in the estimated coil-sensitivity profiles as well as numerical instabilities in the reconstruction further reduce image quality [Blaimer et al., 2013]. Furthermore, the time required for the diffusion encoding is bounded by the achievable and safe maximum gradient strength and slew rate, which limits the achievable in-plane

acceleration [Setsompop et al., 2012].

Advances in parallel imaging have introduced techniques that allow the simultaneous excitation and acquisition of multiple slices using multi-coil arrays [Larkman et al., 2001; Barth et al., 2016]. Using the coil sensitivity profiles, it is possible to reconstruct multiple slices simultaneously, which reduces the scan time of a volume by a factor equal to the number of simultaneously excited slices with a moderate SNR penalty. Blipped gradient pulses can be used to control the aliasing of the excited slices by shifting the slices in the phase encode direction, further increasing the signal to noise and achievable acceleration [Setsompop et al., 2012].

3.2.4. Artefacts

EPI requires a comparatively long signal readout and strong gradients. This has implications for the image quality and robustness to magnetic field inhomogeneities and motion and makes diffusion MRI subject to image artefacts [Le Bihan et al., 2006].

3.2.4.1. Susceptibility artefacts

The interface between areas of different magnetic susceptibility such as brain tissue and air filled cavities or bone causes local magnetic field inhomogeneities. This results in a local modulation of the spins' precession frequency, which induces a local shift in position along the phase encoding direction, and local compression or stretching due to the variations in these shifts in the phase encode direction, leading to image distortions.

The time between frequency encode samples in EPI ("dwell time", about $5\mu s$) is limited by the receiver bandwidth (the frequency range that can be sampled in that time). However, at typical image resolutions, the time between samples in the phase encode direction is two orders of magnitude longer than in the frequency encode direction; or equivalently the effective bandwidth of the phase encode direction is much smaller than that along the frequency encode direction.

Susceptibility differences in the head, especially close to the fontal cortex and the medial temporal lobe, can cause shifts of several voxels along the phase encode direction, resulting in strong distortion with hyper- and hypointense image areas due to compression and expansion, respectively. These distortions can be corrected using additional data to estimate the magnetic field inhomogeneities (field maps), which can be used to unwarp the image distortions. However, areas where voxels were compressed have lost spatial information. This loss of information can be alleviated by non-linear registration and combination of images acquired with different phase encode directions [Andersson, Skare, Ashburner, 2003].

3.2.4.2. Eddy currents

The strong fast-switching gradients in EPI cause eddy (Foucault) currents in conducting material in the scanner. These current loops counteract the change of the magnetic field caused mainly by the diffusion-weighting gradients at high b-values but also by the imaging gradients (Lenz's law). Eddy currents introduce a time-delay and reduced

amplitude of those time-varying gradients and can be present long after the gradients were switched off.

Eddy currents present during readout can for instance cause misalignment in the k-space trajectory of the readout. If they produce a magnetic field component along the frequency encode direction, they cause the k-space lines to drift, resulting in a sheared image. Eddy current fields in the phase encode direction alter the k-space line spacing, which results in a stretched or compressed image in the y-direction. Fields along the B_0 axis introduce a frequency shift and hence a shift in the frequency encode direction that is dependent on the slice location along the B_0 axis.

It is possible to reduce the effect eddy currents have on the net magnetic field by modifying the waveform fed into the amplifier to compensate for the expected effects of eddy-currents ("pre-emphasis", "pre-compensation") but this does not prevent eddy currents. Actively shielded gradient coils are designed to minimise the field outside the coil, and so to minimise the production of eddy-currents in any materials outside the coil [Hidalgo-Tobon, 2010]. The remaining distortions can be further corrected using image registration techniques [Haselgrove, Moore, 1996].

Diffusion analysis techniques typically require multiple images acquired with different diffusion encoding directions and strengths. However, those images would be affected by different eddy currents and therefore different distortions. Hence, accurate spatial alignment requires post-processing techniques that correct for susceptibility artefacts, eddy current artefacts and bulk motion between volumes [Andersson, Sotiropoulos, 2015b].

3.2.4.3. Bulk motion

Bulk or rigid body motion of the head in in-vivo diffusion MRI causes phase errors in the transverse signal that manifest in image artefacts. If the brain can be treated as a rigid body and if the motion is small enough to be negligible in the absence of the diffusion weighting gradient, then this motion can be decomposed into a translation and a rotation. Each have different effects on the magnitude and phase of the acquired data. A small translation causes a phase shift in the signal but does not affect the magnitude image [Hahn, 1960]. A rotation, on the other hand, introduces phase gradients that depend on the direction of the rotation axis and that of the diffusion weighting gradient [Anderson, Gore, 1994; Trouard et al., 1996]. If uncorrected, these phase gradients cause blurring, signal dropout and image distortions in multi-shot EPI that can be corrected fully or to some degree using additional 1D, 2D or 3D navigator echoes. This is under the assumption that the head moves in a trajectory that is coherent while the EPI sequence is performed and that the head performs a rigid body motion [Norris, 2001].

Even though single-shot EPI is relatively robust to bulk motion, acquisition of volumes covering the brain, using multiple b-values and diffusion encoding directions, requires spatially alignment to be consistent across volumes and across slices within each volume. This retrospective alignment is commonly performed using image registration [Anderson, Gore, 1994] and reorientation of the diffusion gradient directions [Leemans, Jones, 2009].

3.2.4.4. Pulsatile artefacts

The arterial vasculature in the brain expand and shrink with the cardiac cycle. This pulsatile motion affects also the brain, which moves with velocities up to $1.5mm/s$ in the vicinity of vessels and in spatially varying temporally nonlinear patterns [Greitz et al., 1992]. During a cardiac cycle, the arterial, venous and capillary blood volumes change in complex synchrony with the brain and CSF volume.

In the presence of diffusion weighting gradients, cardiac pulsation in parenchyma causes incoherent motion on a voxel-scale, which introduces nonlinear phase shifts and therefore signal dropout [Poncelet et al., 1992]. Cardiac gating can be used to acquire data outside pressure peaks (diastole) and to make the signal more consistent between repetitions [Brockstedt et al., 1999] compared to non-gated EPI, but has the drawback that it prolongs the total acquisition time and introduces irregular waiting times in the sequence, which, if a short TR is used, can introduce spin-history artefacts as discussed in the next section.

3.2.4.5. Cross talk and spin-history

The imaging sequence changes the state of the spins in the sample. These perturbations persist when the repetition time (TR) is shorter than the spin relaxation. The T_1 relaxation time in brain tissue is about 0.8s to 1.3s [Wansapura et al., 1999] and in CSF about 3s [Condon et al., 1987]. Therefore, at a TR of 9s, more than 95% of the signal should be in the original state. Imaging repeatedly at shorter (but constant) TR can bring the system into a steady-state. However, this poses the risk of spatially varying degrees of recent exposure to previous pulses ("spin history"), which induces spatially varying intensity modulations. These spin-history effects can be caused by subject movement or temporal inconsistencies between slices or volumes, introduced by cardiac gating or variable length calibration scans.

Furthermore, slice selection pulses are not perfectly constrained to the imaged slice due to hardware limitations. It is also desirable to use excitation profiles that are slightly wider than the slice thickness to be robust to motion relative to the slice encoding direction. However, since slices are acquired consecutively, each excitation would interference with the state of the slice to be imaged next. This "cross-talk" can be reduced by acquiring slices in an interleaved fashion where even and odd slices are acquired separately. However, in the presence of spin-history artefacts, this temporally interleaved acquisition pattern can cause sharp stripe patterns along the slice direction.

3.3. Biophysical correlates of diffusion measurements

Moseley [Moseley et al., 1990] observed that at image resolutions of $3mm$, diffusion in white matter is anisotropic but isotropic in grey matter. In axon bundles or muscle fibres, the diffusion signal is less attenuated in the direction perpendicular to the fibre axes than along the fibres. These findings suggest that diffusion in biological tissue is hindered or

restricted by structures on the scale of the size of cells.[2] This statistical coincidence allows diffusion MRI to probe microstructural properties of the tissue, despite the image resolution being on the order of mm and has made diffusion-weighted MRI an established tool to study tissue microstructure and the organisation of cells across the whole brain in vivo.

Diffusion contrast allows measuring differences of the diffusion characteristics across areas in the brain, between subjects and over a period of time. Yet, inferring from the signal what actually caused a difference is an inverse problem that requires detailed knowledge about microscopic properties of brain tissue [Edgar, Griffiths, 2009] and their effect on the signal given a specific diffusion sequence.

Possible cellular processes that contribute to the diffusion contrast are osmosis due to concentration gradients, active transport along the highly organised microtubule and active or passive transport across cell membranes.

3.3.1. Cell membranes and myelin

The most likely largest contribution to the diffusion contrast are cell membranes. Beaulieu, Allen showed that even unmyelinated axons and axons with depolymerised microtubule exhibit strong diffusion anisotropy [Beaulieu, Allen, 1994].

Cell membranes are composed of lipid bilayers consisting of a hydrophobic interior. This reduces the rate at which polar molecules such as water can diffuse across the membranes [Verkman et al., 1996] and contributes in a major form to the observed diffusion signal by restriction of water and hindrance through permeable membranes or as a net lower mean squared displacement due to densely packed membranes. Cell membranes are a major factor of the observed diffusion anisotropy in neural tissue [Beaulieu, 2002].

Within membranes, proteins and therefore protein-bound protons perform anisotropic diffusion [Saffman, Delbrück, 1975]. Yet this movement is invisible on typical diffusion weighting time-scales due to the very short T_2 on the order of tens of μs [Deese et al., 1982; Wilhelm et al., 2012]. Membranes also affect water molecules in their direct vicinity through chemical and physical processes [Deese et al., 1982; Stanisz et al., 2005], which is the basis of T_1, T_2 and magnetisation transfer imaging [Stanisz et al., 2005; Stanisz et al., 1999; Henkelman, Stanisz, Graham, 2001] and these effects, markedly the T_2 weighting, are superimposed on the diffusion weighting of the signal.

Myelin has a relatively short T_2 of between $10ms$ and $40ms$ [Mackay et al., 1994; Mackay et al., 2006]. This greatly attenuates the contribution of those protons in most diffusion imaging sequences [Wilhelm et al., 2012; Barkovich, 2000]. However, anisotropic diffusion of myelin-associated water has been shown using diffusion sequences with T_1- or T_2-selective excitation pulses [Andrews, Osborne, Does, 2006].

On a larger length-scale, cell membranes form complex landscapes of barriers of various permeability and chemical environment. Electron micrographs offer fascinating snapshots

[2] Hindered diffusion in physics refers to the reduced average displacement compared to free diffusion due to collisions with impermeable boundaries on the time-scale of the experiment. These boundaries do not confine particles, merely hinder their movement. In restricted diffusion, particles are surrounded by boundaries, limiting their movement to the volume enclosed by the boundaries [White et al., 2014].

of this mesostructure (see figs. 2.8 and 3.4). Despite local tissue compression and scale distortions of up to 30% introduced by the tissue preparation [Denninger et al., 2014], electron micrographs offer unique 2D or stacked 2D images covering less than an MRI voxel up to whole brain coverage, which allows high-resolution nerve-tracking [Mikula, Binding, Denk, 2012; Mikula, Denk, 2015] and validation of microstructure diffusion model parameters [Mollink et al., 2017; Stikov et al., 2015b].

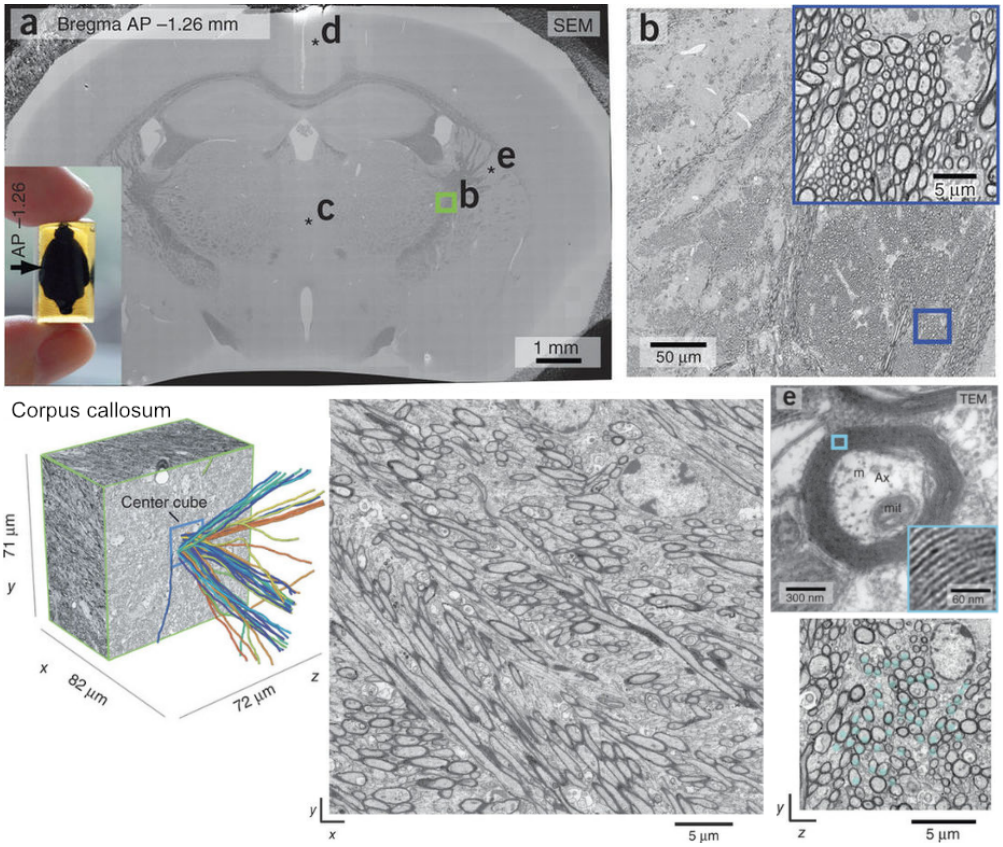


Figure 3.4.: "Block-face image of a whole [mouse]-brain cross-section cut at the level of bregma -1.26, coated with 5 nm platinum-carbon and imaged with scanning electron microscopy (SEM) using secondary-electron detection. Inset, horizontal view of the entire mouse brain after embedding; block dimensions are 6 × 8.5 × 14 mm3. AP, anterior-posterior. (b) Single image tile (green box in a). Inset, magnified subregion indicated by the blue box. ... (e) High-magnification transmission electron microscopy (TEM) image of a 70-nm section taken from a region in the dorsolateral striatum (labeled asterisk in a) in a different sample. Note the intra-period and major dense lines in the inset in e. mit, mitochondrion; Ax, axon; m, microtubules." Bottom left: "Serial block-face electron microscopy stack from the corpus callosum, cut down the middle, with 50 traced axons emerging, randomly colored." Bottom middle and right: individual xy and yz slices of the same block with seed location shown in cyan. Notice the orientation dispersion evident in the 3D rendering and the spatial heterogeneity of fibre alignment in the corpus callosum evident in the cross-sectional image. Adapted from [Mikula, Binding, Denk, 2012] with permission.

3.3.2. Tissue compartments and exchange

Assuming no significant water exchange between the intra and extra-cellular space during a diffusion MRI experiment, molecules residing inside cells are restricted and hindered

in their movement by the cellular boundaries. Similarly, water in the extracellular space is hindered by those membranes, yet, given enough time, molecules can diffuse for long distances in any direction. These two compartments have distinct diffusive characteristics that contribute in an additive fashion to the diffusion signal.

However, biological tissue is very heterogeneous in terms of cell types, sizes and shapes and their arrangements. For instance, the optic nerve in adult male guinea pigs is composed of the following tissue volume fractions (electron microscopic measurements prepared to avoid tissue shrinkage): 32% axoplasm, 25% myelin, 12% astrocyte processes, 16% astrocyte somas, 8% oligodendrocyte processes, 5% oligodendrocyte somas [Perge et al., 2009]. Surprisingly, only 57% by pure volume of the optic nerve contains axons (all axons were myelinated) and astrocytes occupy 14 times the volume of the extracellular space (2%).

Even in coherent appearing tracts such as the central corpus callosum, axons do not follow parallel trajectories (see fig. 3.4) but are inherently disperse [Mollink et al., 2017]. Also, on a coarser scale, white matter bundles intermix and at clinical resolutions, the prevalence of multiple fibre orientations ("crossing fibres") in a voxel is non-negligible with estimates of at least 30% affected voxels in white matter [Behrens, Berg, Jbabdi, 2007; Jeurissen et al., 2013].

Even in the absence of complex geometric arrangements, separating those cellular details from the diffusion signal is challenging. Perge et al. found that individual axons in the optic nerve vary in diameter by a factor of 2 and that 95% of the axons have diameters between $500nm$ and $1.5\mu m$, which cannot be resolved with current diffusion MRI sequences on scanners with gradient strengths of up to $300mT/m$ [Nilsson et al., 2017]. Also, the validity of the assumption that we can neglect exchange between cellular compartments in diffusion MRI of healthy brain tissue is a matter of ongoing debate [Novikov et al., 2016].

The rate at which water crosses cell membranes is subject to complex molecular kinetics [Amiry-Moghaddam, Ottersen, 2003; Stein, 2012]. The cellular water content and ion concentrations are regulated via active transport of water through water channels; aquaporin in astrocytes [Papadopoulos, Verkman, 2013] and possibly other active transport mechanisms in neurons [Yang et al., 2017]. The biological mechanisms that influence the water residence time in neural tissue are not fully understood [Yang et al., 2017].

A fast exchange between the intra- and extracellular space would increase the RMS displacement perpendicular to axons, which would make axons appear less anisotropic or higher calibre if not accounted for [Nilsson et al., 2013b]. The average residence time of water in cells, the intracellular preexchange lifetime, characterises the time taken for 63% of the intracellular water to exchange with the extracellular space and has to be longer than the extracellular preexchange lifetime to maintain homoeostasis in tissue with less extra- than intracellular space [Quirk et al., 2003].

Reported exchange times in brain tissue vary widely from 25ms to 2.5s [Nilsson et al., 2009; Pfeuffer et al., 1998; Quirk et al., 2003; Pfeuffer, Provencher, Gruetter, 1999; Nilsson et al., 2013a] and are heterogeneous across the brain [Lampinen et al., 2017]. In myelinated axons, water residence time increases with myelin thickness [Dortch et al., 2013; Harkins, Dula, Does, 2012]. Astrocytes have higher permeability than neurons

[Solenov et al., 2004; Borgnia et al., 1999; Quirk et al., 2003] and their relatively rapid exchange of water with the extracellular space has lead researchers to build diffusion models that include those glial cells in the "extracellular" compartment [Jespersen et al., 2007; Yablonskiy, Sukstanskii, 2010]. For a review on the role of water exchange on diffusion microstructure estimation see [Nilsson et al., 2013b].

3.4. Diffusion signal representations

The diffusivity of free water at 37.5° is $3.1\mu m^2/ms$ but the measured diffusivity in brain tissue lies between 0.6 and $1.0\mu m^2/ms$, corresponding to a RMS displacement of $11\mu m$ to $14\mu m$ during a diffusion time of duration $100ms$. The measured diffusivity depends on the tissue and is affected by pathologies such as stroke [Warach et al., 1992], which makes diffusion weighted imaging a valuable tool in clinical practice. To distinguish the free diffusion constant D from its estimated counterpart as measured via diffusion-weighted MRI, the latter is commonly referred to as Apparent Diffusion Coefficient (ADC) [Le Bihan et al., 1986].

ADC summarises all tissue properties in a single value. Diffusion Spectrum Imaging (DSI) [Wedeen et al., 2005] lies on the other extreme of signal representations. We assumed free diffusion in the absence of structure influencing the molecules in the derivation for equation 3.23, which describes the signal decaying in a mono-exponential way. However, one can use the general equation 3.21 to reconstruct the conditional probability distribution $\mathcal{P}(\mathbf{x}, t|\mathbf{x}_0, 0)$ by means of Fourier transformation of the diffusion signal measured with q-vectors with different diffusion sensitisation strengths and directions, spanning the range of interest. DSI gives access to orientation and length-scale resolved tissue properties. The drawback of this method, however, is that it requires strong gradients and long acquisitions to sample q-space with sufficient density and coverage, which limit its clinical applicability [Lätt et al., 2007a; Lätt et al., 2007b].

Therefore, the most common techniques acquire the signal for multiple diffusion encoding directions to sample the orientation dependence of the diffusion attenuation but with a constant gradient weighting strength ($||\mathbf{q}|| = \text{const}$). This is referred to as sampling on a single b-value "shell". In the following sub-sections, I will briefly discuss signal representations that use data acquired on a single or on multiple shells to capture information about the tissue.

3.4.1. Diffusion tensor

Diffusion Tensor Imaging (DTI) [Basser, Mattiello, LeBihan, 1994] is a method that aims at measuring the diffusion anisotropy in the tissue under the assumption that the probability distribution $\mathcal{P}$ can be approximated as a multivariate Gaussian distribution

$$\mathcal{P}(\mathbf{x}, t|\mathbf{x}_0, 0) = \frac{1}{(4\pi t|\underline{D}|)^{3/2}} e^{\frac{(\mathbf{x}-\mathbf{x}_0)^T \underline{D}^{-1}(\mathbf{x}-\mathbf{x}_0)}{4t}} \tag{3.29}$$

with the apparent diffusion coefficient replaced by the apparent diffusion tensor

$$\underline{D} = \frac{1}{6t}\left\langle (\mathbf{x}-\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)^T \right\rangle = \begin{bmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{bmatrix} \tag{3.30}$$

This is a generalisation of the mean-square particle displacement being uniformly defined by the variance of a Gaussian distribution (see eqs. (3.6) and (3.13)) to the arbitrarily-oriented multivariate Gaussian distribution in eq. (3.29).

$\underline{D}$ is a symmetric ($D_{ij} = D_{ji}$), positive-definite[3], rank-2 tensor (matrix) [Basser, Mattiello, LeBihan, 1994]. Or put in physical terms, the mean square distance has to be larger than zero and the probability distribution is point-symmetric with respect to $\mathbf{x}_0$ ($\mathcal{P}(\mathbf{x}, t|\mathbf{x}_0, 0) = \mathcal{P}(\mathbf{x}_0, t|\mathbf{x}, 0)$).

For gradient direction $\mathbf{e}_G$, the corresponding apparent diffusivity is the projection of $\underline{D}$: $\mathbf{e}_G^T \underline{D} \mathbf{e}_G$. The signal attenuation in direction $\mathbf{e}_G$ in the diffusion tensor model is therefore

$$S(b, \mathbf{e}_G) = S(b=0) e^{-b\mathbf{e}_G^T \underline{D} \mathbf{e}_G} \tag{3.31}$$

Or expressed as a matrix product $S(b) = S(b=0)e^{-\sum_{i,j} b_{ij} D_{ij}}$, the scalar b (eq. (3.25)) is replaced by a 3x3 matrix b_{ij} defined as [Yablonskiy, Sukstanskii, 2010]

$$b_{ij} = \gamma^2 \int_0^{TE} \left(\int_0^t \mathbf{G}_i(t')dt' \int_0^t \mathbf{G}_j(t')dt' \right) dt \tag{3.32}$$

By measuring at least six non-collinear diffusion encoding directions on a single shell and an additional typically non-diffusion weighted measurement (S_0), one can determine the diffusion tensor in a voxel by solving eq. (3.31) for the coefficients of the diffusion tensor.

$\underline{D}$ depends on the orientation of the sample with respect to the scanner coordinate system. It is therefore common to report a scanner coordinate system independent representation of the diffusion tensor. Since $\underline{D}$ is a positive definite matrix, we can decompose it via eigenvalue decomposition

$$\underline{D} = \underline{E}\underline{\Lambda}\underline{E}^{-1} = \underline{E}\underline{\Lambda}\underline{E}^T = \underline{E} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \underline{E}^T \tag{3.33}$$

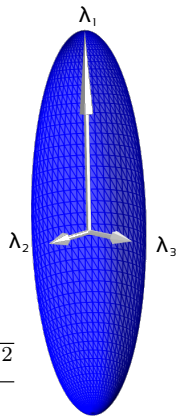
with $\underline{E}$, an orthonormal matrix and Λ a real-valued positive diagonal matrix. The columns of $\underline{E}$ ($\mathbf{E}_i$) are eigenvectors of $\underline{D}$ and the diagonal entries of $\underline{\Lambda}$ are the corresponding eigenvalues: $\underline{D}\mathbf{E}_i = \lambda_i \mathbf{E}_i$ that we can assume to be sorted so that $\lambda_1 >= \lambda_2 >= \lambda_3$.

The matrix $\underline{E}$ defines the direction along which the diffusion tensor is oriented and can be used to extract the principal direction ($\mathbf{E}_1$) of the diffusion tensor. The eigenvalues can be used to derive scalar measures that describe rotation invariant properties of the diffusion tensor [Le Bihan et al., 2001] such as the diffusivity along the principal direction ("axial diffusivity"), the average diffusivity in the directions perpendicular to that ("radial

[3] $\mathbf{x}^T \underline{D} \mathbf{x} > 0$, for every non-zero real-valued vector $\mathbf{x}$

diffusivity"), the average or mean diffusivity, and the degree of anisotropy (Fractional Anisotropy (FA)). FA is normalised to the range between zero for isotropic diffusion and one for anisotropic diffusion (λ_2 and λ_3 are negligible compared to λ_1).

$$\begin{aligned}
\text{axial diffusivity } D_a &= \lambda_1 \\
\text{radial diffusivity } D_r &= \frac{\lambda_2 + \lambda_3}{2} \\
\text{mean diffusivity } D_m &= \frac{\lambda_1 + \lambda_2 + \lambda_3}{3} \\
\text{fractional anisotropy FA} &= \frac{\text{std}((\lambda_1, \lambda_2, \lambda_3))}{\text{RMS}((\lambda_1, \lambda_2, \lambda_3))} \\
&= \sqrt{\frac{3}{2}} \frac{\sqrt{(\lambda_1 - D_M)^2 + (\lambda_2 - D_M)^2 + (\lambda_3 - D_M)^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}
\end{aligned}$$



3.4.2. High Angular Resolution Diffusion Imaging (HARDI)

The diffusion tensor model is not capable of unambiguously representing multiple fibre populations with different orientations or diffusive properties in a voxel [Alexander, Barker, Arridge, 2002]. Partial volume effects of multiple tissue compartments with high anisotropy aligned in different directions for instance cause a net lower fractional anisotropy and the principal eigenvector direction does not coincide with the main fibre directions [Alexander, Seunarine, 2010].

To address this issue, diffusion data can be acquired at higher angular resolution (High Angular Resolution Diffusion Imaging (HARDI)), which allows higher order models to extract directional-resolved microstructural information in each voxel [Tuch, 2002b]. In HARDI, contrary to DSI, the q-space is sampled sparsely in the radial direction [Tuch et al., 2002a]. Single-shell HARDI is acquired with a fixed radial component with directions spread to minimise (antipodal symmetric) density variations in the angular domain. Multi-shell HARDI is an extension that uses data acquired on multiple b-value shells. The single-shell techniques sample the direction-dependency of the diffusion attenuation for a given diffusion weighting strength; the multi-shell version can be used to resolve additional intra-voxel tissue properties [Tuch et al., 2002a; Descoteaux et al., 2006; Assaf, Basser, 2005; Alexander, 2008; Jensen et al., 2005; Jeurissen et al., 2014].

3.5. Diffusion signal models

A voxel in a diffusion MRI image averages the motion of water molecules that are influenced in their movement by a complex biophysical environment. The question is, which processes lead to changes in the observed signal and how to infer the latter from the former. This has since its conception [Stejskal, Tanner, 1965] remained an active field of research with many open questions [Novikov et al., 2016; Novikov, Kiselev, Jespersen,

2018].

3.5.1. Tissue properties and length scales

The mm-scale spatial resolution of MRI allows capturing a detailed map of the whole brain, which can be used to investigate its macrostructure including gyrification, surface and volume measures and anatomical features on the scale of white matter bundles. Diffusion MRI with diffusion times of $1ms$ to $1s$ probes molecular motion at much smaller length-scales of $1\mu m$ to $45\mu m$ and is sensitive to properties of the intra and extracellular environment and, on larger length scales, to the spatial arrangement and alignment of cells. The former is usually referred to as the tissue microstructure and the latter as tissue mesostructure [Novikov et al., 2016; Reisert et al., 2017].

The microstructural fingerprint of a tissue compartment depends on the μm-level T_1 and T_2 relaxation rates and deviation from the Gaussian diffusion profile it causes on the observed (averaged) signal [Novikov et al., 2016]. It captures the molecular neuroanatomy [Pollock, Wu, Satterlee, 2014] and biochemistry of the brain [Kinney et al., 1994].

The mesostructure can be thought of as the coarse alignment of the microstructurally distinct compartments in the voxel and expresses the orientationally-resolved compartment volume fractions and is influenced by the fibre orientation distribution (spherical probability distribution of fibre directions), fibre crossings, fanning, bending and undulation. Different diffusion models use different measures and terms when they refer to the mesostructure: NODDI: "fibre dispersion" [Zhang et al., 2012], CSD: "fibre orientation distributions" [Tournier, Calamante, Connelly, 2007]. Separating microstructural properties to derive tissue-specific mesostructure is a difficult, potentially degenerate modelling problem [Novikov et al., 2016] and different models (or fitting procedures) have to make - possibly tissue-specific [Novikov et al., 2016] - assumptions about model parameters [Reisert et al., 2017].

By connecting the voxel-wise micro and mesoscopic information, diffusion MRI allows to derive unique macroscopic properties that can be used to infer long range connections in the data via tracking techniques ("tractography") [Basser et al., 2000; Mori et al., 1999]. Resulting tractograms depend on the extracted microstructural information and on the tracking algorithm, each with their assumptions and limitations. Unfortunately, the uncertainty and ambiguity associated with fibre tracking does not warrant an interpretation of the resulting tractograms as a measure of the true white matter fibre connections and their "strengths" [Donahue et al., 2016; Girard et al., 2014; Jeurissen et al., 2017]. However, the resulting streamlines can be used for the extraction of local or global connectivity patterns [Gong et al., 2008] in the diffusion data and can guide statistical analysis of diffusion properties along the path [Raffelt et al., 2015].

3.5.2. Compartment models

The aim in using tissue compartment models is to find a set of parameters that best match the diffusion data and use these parameters to reason about the mesoscopic properties of the tissue. The step from model parameters to tissue properties naturally depends

on the validity of the model assumptions, the interpretability of the model parameters, the robustness of the fitting procedure to noise and the model behaviour in the presence of biological heterogeneity and pathology. A common assumption is that the contributions to the diffusion weighted signal $S(\mathbf{G}(t))$ can be expressed as a weighted sum of independent compartment contributions

$$S(\mathbf{G}(t)) = S(\mathbf{G} = 0) \sum_c f_c m_c(\mathbf{G}(t), \mathbf{p}_c) \tag{3.34}$$

where f_c is the volume fraction (possibly weighted by the compartment-specific T_2 decay) of compartment c, and m_c the corresponding signal attentuation given the diffusion weighting $\mathbf{G}(t)$ and set of compartment parameters $\mathbf{p}_c$.

Most models of white matter are restricted to 2 to 3 compartments and contain typically a total of 4 to 12 parameters that are fitted to the diffusion signal [Ferizi et al., 2015]. However, the number of parameters might be limited to 10 or 11 by the resolution limit of current PGSE diffusion sequences and scanner hardware and fitting stability decreases with increasing number of parameters [Ferizi et al., 2015]. Despite the relatively low number of parameters, there is a growing variety of microstructure models. Some of the models are reviewed and compared in [Panagiotaki et al., 2012; Ferizi et al., 2015; Jelescu, Budde, 2017].

A common assumption of compartment models is that the intra-axonal diffusivity is nearly unhindered parallel to the axon axis but zero or very small in the transverse direction. A hindered but anisotropic compartment is prevalent to model diffusion in the extracellular space and possibly including contributions from diffusion through glial cells, assuming highly permeable membranes. An isotropic compartment ("ball") is used to model free water or water trapped in cells, depending on the radius of that ball. Depending on the shape of the assumed anisotropic diffusion profile or their parametrisation, anisotropic compartments are referred to in a number of ways: sticks, cylinders, zeppelins or tensors.

Besides neurite orientation and density [Jespersen et al., 2007; Zhang et al., 2012], models are geared to derive different tissue properties such as axon diameter [Assaf et al., 2008], cell permeability [Nedjati-Gilani et al., 2017] and orientation dispersion [Zhang et al., 2012]. The NODDI framework [Zhang et al., 2012] assumes fixed and predefined intra-axonal, extracellular (axial), and isotropic diffusivities and supposes a linear relation between the axial and radial diffusivity of the extra-axonal space, depending only on the intra-axonal fraction. These parameters are used to fit the volume fractions of the intra-axonal and isotropic volume compartments and a dispersion parameter that is modelled by a distribution of zero-diameter fibre bundles (sticks), with angular deviations from the mean direction characterised by a Watson distribution. These choices narrow the free parameter pool, which is sensible from an algorithmic stability point of view but they are problematic for multiple reasons. The model does not take a tissue or sequence dependency of the diffusivities into account. This leaves volume fractions and dispersion as the only free parameters to model changes in diffusivity, rendering the model's parameters unspecific to the actual microstructural properties it explicitly models [Jelescu, Budde, 2017]. NODDI does not model fibre crossings, hence a dispersion index

could be attributed to fibre crossings or fibre dispersion [Kaden et al., 2016] and the relationship between axial and radial extra-axonal diffusivity is an approximation that is violated in Monte Carlo simulations and analytical models for packing densities in biologically plausible ranges [Novikov, Fieremans, 2012].

Overall, despite explaining the diffusion signal well and producing values that lie in plausible ranges, some of the derived parameters can differ significantly between models [Novikov et al., 2016] and there seems to be no clear best model [Ferizi et al., 2015; Reisert et al., 2017; Jelescu, Budde, 2017].

3.5.3. Constrained Spherical Deconvolution

Spherical convolution Let the axially symmetric function $R(b, \theta)$ ("response function") characterise the diffusion signal profile of white matter fibres in the brain as a function of the b-value and azimuth relative to the fibre axis (θ) [Tournier et al., 2004]. In the absence of exchange between fibre populations, the HARDI signal can be expressed as a weighted sum of differently oriented fibre populations

$$S(b, \theta, \phi) = \sum_i f_i A_i \left(R(b, \theta)\right) \tag{3.35}$$

with A_i, the operator that rotates the response function $R(b, \theta)$ from its arbitrary initial direction into the ith fibre population's direction. Going from a discrete weighted sum to an integral over all fibre directions, one can write eq. (3.35) as a spherical convolution of the response function with $f(\theta, \phi)$, the Fibre Orientation Distribution (FOD) [Tournier et al., 2004].

$$S(b, \theta, \phi) = f(\theta, \phi) * R(b, \theta) \tag{3.36}$$

Tissue response function The white matter response function can be simulated using any of the anisotropic and isotropic model components discussed in section 3.5.2. However, it can also be sampled from the data; more precisely from voxels that appear to contain only white matter and whose white matter orientation distribution is sharp (high FA) so that one can assume that the voxel contains a single collinear fibre population. Note that white matter voxels contain mostly non-axonal tissue and extracellular matrix (compare section 3.3.2) and that the white matter response function is not specific to the angular and b-value dependency of the diffusion signal of parallel axons. The white matter response function is assumed to capture the diffusion profile of a voxel containing a representative sample of a single strand of compact, coherently aligned white matter —including associated non-axonal cells and extracellular matrix. After voxel-wise reorientation of this "single fibre" signal to a common direction, the signal can be averaged to obtain the white matter specific response function [Tournier, Calamante, Connelly, 2007].

For typical clinical diffusion gradient durations and b-values larger than $3000 s/mm^2$, the radial diffusion weighted signal originating from the extracellular volume fraction (and highly permeable cells such as astrocytes, see section 3.3.2) is highly attenuated

in white matter fibres. The radial signal of the restricted intra-axonal compartment, however, is almost completely preserved and approximately proportional to the intra-axonal volume fraction [Raffelt et al., 2012]. Note that the contribution to the signal from restricted small cells ("dot" compartment) is likely negligible in white matter [Tax et al., 2018]. As the FOD amplitude is approximately proportional to the radial signal, Raffelt et al. coined the term Apparent Fibre Density (AFD) for the FOD amplitude to indicate that the FOD under these imaging conditions is sensitive to the "intra-axonal volume fraction of the axons running along the corresponding orientation" [Raffelt et al., 2012]. The specificity of AFD to the fibre intra-axonal volume fraction depends on the b-value and on the degree to which contributions from other cell types can be ignored.

The presence of multiple tissue types that exhibit different diffusion profiles, for instance due to differences in relaxation rate, water exchange, or packing density, are not explicitly modelled in constrained spherical deconvolution (CSD). Hence deviations from the white matter response function would be captured by an altered FOD amplitude. This makes CSD sensitive to pathology [Raffelt et al., 2012] and places CSD somewhere between a pure signal representation that has no interpretable value and the microstructural tissue models that explicitly model and fit tissue micro- and mesostructural properties.

Spherical Harmonics A Fourier series allows representing a continuously differentiable function on the unit circle as a uniformly convergent series of weighted sinusoids and cosines with increasing frequency. Equivalently, a Laplace series can be used to represent a function in $\mathcal{R}^3$ that is continuously differentiable on the unit sphere as a uniformly convergent series of spherical harmonics [Kalf, 1995].

The decomposition of the function $f(\theta, \phi)$, which is defined on the surface of a sphere is

$$f(\theta,\phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} c_l^m Y_l^m \tag{3.37}$$

with c_l^m the weighting coefficients, and Y_l^m the spherical harmonic basis functions of order l.

Using the normalisation $\int_0^{2\pi} \int_0^{\pi} Y_l^m \hat{Y}_{l'}^{m'} sin\theta d\phi d\theta = \delta_{ll'}\delta_{mm'}$, with $\hat{Y}_l^m$, the complex conjugate of Y_l^m and $\delta_{mm'}$, the Kronecker delta, Y_l^m becomes [Descoteaux et al., 2006]

$$Y_l^m = \sqrt{\frac{(2l+1)}{4\pi}} \sqrt{\frac{(l-m)!}{(l+m)!}} P_l^m\left(\cos(\theta)\right) e^{im\phi} \tag{3.38}$$

with $\phi \in [0, \pi]$, $\phi \in [0, 2\pi]$ and $P_l^m(\xi)$, the associated Legendre polynomial of order $l \geq 0$ and degree $|m| \leq 0$.

When diffusion signal and fibre response are represented in spherical harmonics, the convolution of the response function with the FOD (eq. (3.36)) becomes a set of linear matrix vector multiplications, in analogy to the convolution theorem in the Fourier series [Healy Jr, Hendriks, Kim, 1998]. This allows the efficient estimation of the FOD using

the known fibre response function $R(\mathbf{G}(t), \theta)$ and measured HARDI signal [Tournier et al., 2004].

In analogy to the frequency in the Fourier series, spherical harmonics of higher order can represent increasingly sharper functions. The uniformly convergent property of the series ensures that the deviation from the function becomes smaller with increasing order, which allows smooth functions to be approximated using a truncated series of spherical harmonics up to a finite order l_{max}. This allows representing the HARDI signal by projection onto the spherical harmonic basis up to a predefined order [Frank, 2002].

The $l = 0$ term ("DC"-term) is the average value of the function over the sphere [Sloan, 2008] and the higher order terms capture the angular frequency component. If the signal (or the response function) is aligned with the z-axis, it follows from the axial symmetry that all $m \neq 0$ phase terms are zero. The series of pure m=0 terms are also referred to as zonal spherical harmonics.

Note that the change of basis to spherical harmonics in itself is not very useful as it is simply a different representation of the signal lacking biological interpretability. However, spherical harmonics are computationally advantageous in decomposing the signal into an orientation-resolved density ("orientation distribution"), given a kernel that characterises the diffusion signal characteristics in a given direction [Anderson, 2005; Tournier et al., 2004]. The uniformly convergent property of spherical harmonics is convenient for analysing the FOD expressed in spherical harmonics as it gives direct access to the average contribution of the white matter signal in the voxels (DC term). Tissue configurations with close to uniform angular profiles on the voxel level (grey matter) have small coefficients in the higher order terms.

Non-negativity constraint In practice, spherical deconvolution involves an ill-conditioned matrix inversion that results in unstable FOD estimates that yield unphysical negative FOD amplitudes. This can be addressed with a non-negativity constraint on the FOD amplitude[Cheng et al., 2014; Tournier, Calamante, Connelly, 2007]. The FOD can be estimated as the following constrained linear least squares problem:

$$\arg\min_{\mathbf{f}} \frac{1}{2} \|\underline{C}_b \mathbf{f} - \mathbf{S}_b\|_2^2, \text{with } \underline{A}\mathbf{f} \geq \mathbf{0} \tag{3.39}$$

with $\mathbf{S}_b = [S(b, \theta_1, \phi_1), ..., S(b, \theta_N, \phi_N)]^T$ the HARDI signal intensities of shell b measured (or super-resolved) in N directions concatenated to a vector, $\mathbf{f}$ the vector of FOD indices in the spherical harmonic basis, $\underline{C}$ the matrix that performs the convolution on the coefficients of $\mathbf{f}$, and $\underline{A}$ the matrix that transforms the coefficients of $\mathbf{f}$ to signal amplitudes [Tournier, Calamante, Connelly, 2007].

3.5.4. Multi-Shell Multi-Tissue Constrained Spherical Deconvolution

The FOD captures the orientation-resolved density of the signal attributed to white matter. However, estimating the FOD in brain regions containing contributions of cerebrospinal fluid (CSF) and grey matter (GM) leads to noisy FOD estimates [Dell'Acqua et al., 2010; Roine et al., 2014] and looses its meaning in non-white matter voxels. To

account for partial volume effects, Jeurissen et al. proposed multi-shell multi-tissue constrained spherical deconvolution (MSMT-CSD), which uses multi-shell HARDI data to deconvolve the signal into n distinct tissue types defined by their respective b-value dependent response function $R_t(\mathbf{b}, \theta, \phi)$ with $t = 1, ..., n$. Extending eq. (3.39) to n tissue types and HARDI signals acquired on m shells, this yields the optimisation problem

$$\arg\min_{\mathbf{f}_t} \frac{1}{2} \left\| \underline{\underline{C}}_{\mathbf{b}t} \mathbf{f}_t - \mathbf{S}_\mathbf{b} \right\|_2^2, \text{with } \underline{A}_t \mathbf{f}_t \geq \mathbf{0} \tag{3.40}$$

Here $\mathbf{S}_\mathbf{b}$ denotes the concatenation of the m shell-specific signal vectors ($\mathbf{S}_b$), $\mathbf{f}_t$ the concatenation of the n tissue-specific orientation distribution function (ODF) and each entry in the tensor $\underline{\underline{C}}_{\mathbf{b}t}$ the shell (rows) and tissue (column) specific convolution operation matrix.

The prerequisite to solve this constrained least squares fitting problem uniquely is that the HARDI data are sampled on at least n shells and that the individual tissue response functions have distinct b-value dependencies [Jeurissen et al., 2014]. In adults, tissue-specific response functions for cortical grey matter and CSF can be derived from coregistered segmented T_1 images [Jeurissen et al., 2014]. Alternative methods derive tissue specific voxel masks by segmentation of the diffusion signal via sparsity-constrained [Jeurissen, Tournier, Sijbers, 2015], or convexity constrained [Christiaens et al., 2015] non-negative matrix factorization or heuristics [Dhollander, Raffelt, Connelly, 2016] based on threshold masking [Ridgway et al., 2009] of the b-value dependency of the direction averaged signal. In contrast to deep grey matter, the diffusion signal characteristics of cortical grey matter in adults are different from that of white matter and CSF, facilitating a decomposition of the signal that matches expected maps of these tissue types [Jeurissen et al., 2014]. In neonates the decomposition of HARDI signals into white matter (WM), GM and CSF is an open research question as grey matter in neonatal HARDI data has signal characteristics very similar to that in deep white matter.

In the presence of local brain pathology such as tumours or lesions, it is possible to represent abnormal tissue as an additional tissue contrast if it has a diffusion signal characteristic that is sufficiently distinct from that of normal brain tissue [Christiaens et al., 2015]. Alternatively, the signal in pathology can be represented using the response functions derived from normal tissue. Abnormality would then manifest as altered tissue density in the resulting orientation distribution functions (ODFs) compared to healthy tissue.

3.6. Conclusion

Diffusion weighted MRI (dMRI) is a non-invasive modality that offers unique insights into tissue properties at the cellular level and beyond, which makes it a valuable tool to study and assess normal and abnormal brain tissue characteristics and to capture the rapidly changing organisation and composition of brain tissue in the perinatal period.

Reliably inferring tissue properties from diffusion data requires knowledge about biophysical properties of tissue, a model how this manifests in diffusion data and a way

of retrieving this information. Although, animal and post-mortem studies have contributed much knowledge about the cellular changes occurring during brain development (see chapter 2), it is an open research question how to characterise these changes using dMRI.

Given the difficulty of most diffusion models to reliably derive tissue characteristics in the mature brain, MSMT-CSD can be an alternative path to capture and characterise tissue maturation, provided it is possible to find meaningful and sufficiently distinct tissue response functions in the data. Under this assumption, MSMT-CSD can be used to reconstruct spatial and temporal maps about brain maturation with directional information about fibrous structures in the brain.

Chapter 4

Classification using Convolutional Neural Networks

Contents

4.1. Introduction

Deep learning is one of the most promising and rapidly evolving[1] fields of machine learning and has transformed computer vision [Goodfellow, Bengio, Courville, 2016]. Deep learning has its name from algorithms that allow the extraction of hierarchical representations of knowledge from data. Especially for computer vision tasks, neuroscience has inspired algorithms for extracting and processing feature representations [Hassabis et al., 2017; Bengio, 2011].

The algorithms for successfully training supervised deep neural networks have been available for decades but their success was limited by the amount of available data and

[1] On average, 400 papers were published each month in 2017 in the computer vision category on arxiv, which is one of the fastest growing categories in computer science https://arxiv.org/help/stats/2017_by_area/index.

sufficient computing capabilities. Specialised hardware (GPUs, TPUs) and large scale data collection efforts [Sun et al., 2017] have facilitated building learning algorithms that for the first time surpass human pattern recognition performance in specific tasks [Schmidhuber, 2015].

The ImageNet classification competition in 2012 showed the capabilities of deep convolutional neural networks in large-scale image classification: the AlexNet [Krizhevsky, Sutskever, Hinton, 2012] architecture beat the second best ranked algorithm by 41%. This caused a shift in the field and in subsequent years, deep neural networks dominated the rankings [Russakovsky et al., 2015].

This chapter gives a brief introduction to image classification using a Convolutional Neural Network (CNN) focusing on aspects most relevant for medical image classification. See [Litjens et al., 2017; Suzuki, 2017; Lee et al., 2017; Bejnordi et al., 2017] for reviews of deep learning in medical image analysis, and recent introductions and reviews of the field of deep learning can be found in [Goodfellow, Bengio, Courville, 2016; Schmidhuber, 2015; LeCun, Bengio, Hinton, 2015].

4.2. Supervised learning: classification

Machine learning algorithms can be separated into 3 categories: supervised, unsupervised, and reinforcement learning. In supervised machine learning, an algorithm is trained to perform a task of learning the mapping from some training data to corresponding "correct" answers that is was provided with, with the goal of generalising this mapping to data outside the training data. An example of supervised learning is linear regression.

Unsupervised learning extracts patterns directly from the data, such as in k-means clustering. In reinforcement learning [Sutton, Barto, 1998], the algorithm is not provided with pre-defined answers but can interact with its simulated environment and learns through reward signals.

In classification, the aim is to discriminate between a defined set of k categorical classes. Presented with some data, a classification algorithm is a function that maps this data to the corresponding category or to a discrete probability distribution representing the likelihood[2] that the sample belongs to either category [Goodfellow, Bengio, Courville, 2016, chapter 5]. Figure 4.1 shows an example of a simulated 2D dataset consisting of two categories (red and blue) and the learned feature space mapping for 3 different classification models. For an introduction to supervised learning and classification see [Robert, 2014; Hastie, Tibshirani, Friedman, 2009a].

[2]Calculating class-correspondence likelihoods without assigning class labels is a regression method, not classification, but this distinction is often not made in machine learning practice.

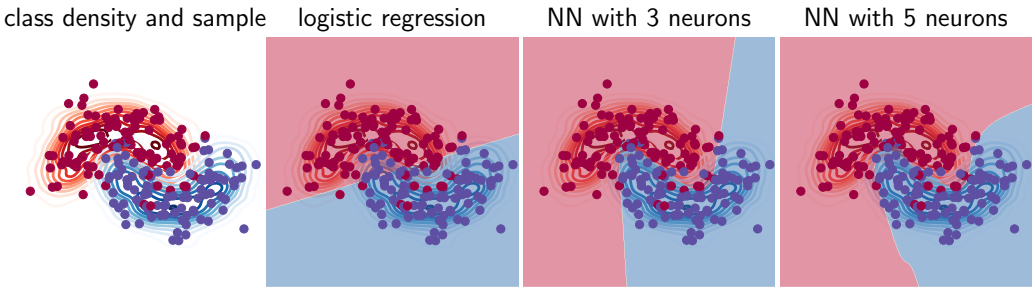


Figure 4.1.: Classification example: Two sampling probability distributions in the shape of half-moons generate samples (dots) of the blue and red class (left). The background colour in the images on the right show the learned mapping of the 2D input space to the category labels for logistic regression and for two neural networks with 3 and 5 "neurons" or units, respectively.

4.2.1. Logistic regression

Binary logistic regression estimates the likelihood that a categorical dependent variable y has the value 0 or 1 given a real valued vector $\mathbf{x}' \in \mathbb{R}^n$. Let $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_n)^T$ be the parameter vector of logistic regression and $\mathbf{x} = (1, x_1, x_2, \ldots, x_n)^T$ the feature vector with an added bias term.

The logistic function $\sigma(z) = \frac{e^z}{e^z+1} = \frac{1}{1+e^{-z}}$ is smooth and monotonously increasing with z and maps any real valued z to the interval [0, 1] [Hastie, Tibshirani, Friedman, 2009b, p 119]. Logistic regression uses this function as a "hypothesis" function h_β to represent the probability that the sample $\mathbf{x}$ belongs to class 1

$$p(y = 1|\mathbf{x}; \beta) = h_\beta(\mathbf{x}) = \frac{1}{1 + e^{-\beta^T \mathbf{x}}} \tag{4.1}$$

Using $p(y = 0|\mathbf{x}, \beta) = 1 - p(y = 1|\mathbf{x}, \beta)$ and that y is either 0 or 1, the probability can be expressed jointly for both classes

$$p(y|\mathbf{x}; \beta) = (h_\beta(\mathbf{x}))^y (1 - h_\beta(\mathbf{x}))^{1-y} \tag{4.2}$$

For a set of m independent samples (x_i, y_i), the likelihood of the parameter vector is [Ng, 2012]

$$L(\beta) = \prod_{i=1}^{m} p(y_i|x_i, \beta) \tag{4.3}$$

Optimal parameters are usually found by maximising the log likelihood

$$l(\beta) = \log(L(\beta)) = \sum_{i=1}^{m} y_i \log(h_\beta(\mathbf{x}_i)) + (1 - y_i)\log(1 - h_\beta(\mathbf{x}_i)) \tag{4.4}$$

This is equivalent to minimising the cross-entropy or the Kullback-Leibler divergence [Kullback, Leibler, 1951] of the predicted label distribution and the training sample

distribution. These are natural distance measures for probability distributions but not the only loss function used for binary classification [Akaike, 1998; Buja, Stuetzle, Shen, 2005].

Note that logistic regression works well if the input features correlate with the decision outcome. For instance, in [Mor-Yosef et al., 1990], multivariate logistic regression classification has been used to recommend cesarean delivery using descriptions of the fetus' and the placenta's presentation and maternal risk factors. However, logistic regression performs poorly directly on MRI images as the intensity of individual voxels is most useful in comparison with other voxels.

4.2.2. Gradient-based optimisation

In machine learning, gradient-based optimisation is an iterative approach to solving a set of equations and is typically used when the equations can not be solved directly, either due to resource constraints or if no analytical solution exists. "Gradient" refers to the derivative of the cost function with respect to the parameters of the system. The parameters are optimised by repeatedly making small steps in the direction that decreases the cost $c(\beta, x_i)$ given the current parameters. The parameter update is

$$\beta_{t+1} \leftarrow \beta_t - \epsilon(t)\frac{1}{s}\sum_{i=1}^{s}\frac{\partial c(\beta, x_i)}{\partial \beta} \tag{4.5}$$

with $\epsilon(t)$ the step size at step t and s is the number of samples that contribute to the update.

Assuming the negative log likelihood of the parameter vector (eq. (4.4)) as the cost function ($c(\mathbf{x_i}, \beta) = -l(\beta)$), the contribution to the update of the logistic regression parameters for a single sample is [Ng, 2012]

$$\frac{\partial l(\beta, x_i)}{\partial \beta} = (y_i - h_\beta(x_i))x_i \tag{4.6}$$

Algorithmically, optimising $l(\beta)$ is equivalent to finding β by minimising the mean squared error between the predicted and the training labels. However, due to the non-linear mapping of the sigmoid function, the mean squared loss function for logistic regression is non-convex and slow to optimise [Goodfellow, Bengio, Courville, 2016, chapter 6].

Using the current best estimate for the parameters, eq. (4.5) estimates the parameter update as the direction in which the cost of the training data decreases the quickest. If this estimate is calculated using all available training samples, the algorithm is called batch gradient descent. However, one can also update the parameters after observing a single example ($s = 1$, "online stochastic gradient descent") or by summation of gradients from a small number of samples between 1 and m ("mini-batch stochastic gradient descent"). A batch refers to the set of samples from which an update of the network parameters is calculated and applied. A training epoch is finished when all samples have been used once.

There are a number of different optimisation algorithms that attempt to speed up convergence and improve robustness to local minima. See [Goodfellow, Bengio, Courville, 2016, chapter 8] and [Ruder, 2016] for an overview. All experiments in chapter 6 use the gradient descent optimisation algorithm "Adaptive Moment Estimation" (Adam) [Kingma, Ba, 2014], which is a universal and robust default optimiser for training neural networks [Ruder, 2016]. Adam uses parameter specific learning rates ($\epsilon(\beta_i, t)$) that are adapted during training using exponentially decaying estimates of the mean and the variance of the gradients.

4.3. Convolutional neural networks

Neural networks consist of connected units or "neurons". A neuron performs a weighted addition and summation of its input signals, adds a bias term and optionally transforms this number via an activation function (see fig. 4.2). In feedforward neural networks, data and computation flows in one direction (see fig. 4.3); neurons pass their output on to the the next layer of neurons or finally to the output layer. Weight and bias terms are learned parameters and the number and arrangement of neurons and the types of activation functions are typically fixed.

A very common activation function is the rectified linear unit ("relu") $\text{relu}(x) = \max(0, x)$ and is typically used throughout the network. This allows the network to combine features non-linearly and to loose information due to the surjectivity of the transformation [Geiger, Feldbauer, Kubin, 2011]. Similarly to logistic regression, sigmoid activation functions can be applied to the output of a neural network to obtain a classification model.

Although the basic building blocks of neural networks are relatively simple and well-understood operations, their combination allows building - maybe surprisingly [Anderson, 1972; Gu et al., 2009] - complex systems that can store abstract representations of information distributed throughout the network [Hinton, 1986].

Convolutional neural networks [Fukushima, 1979; LeCun, Bengio, 1995] are a special type of artificial neural networks and are very successful in computer vision, text and speech analysis, domains where the input to the network is high dimensional and has spatially or temporally local patterns that have a meaning that is invariant to some degree of affine transformation of the input [LeCun, Bengio, Hinton, 2015; Schmidhuber, 2015].

Convolutional neural networks use filters that, in analogy to the visual cortex [Eickenberg et al., 2017; Bengio, 2011], are sensitive to certain local patterns such as edges. The filter size is typically very small and is referred to as the receptive field. By increasing the filter size or by stacking multiple layers, convolutional neural networks can learn to extract representations from larger areas of the image and to combine those to more and more complex patterns similar to processes in the human cortex [Riesenhuber, Poggio, 1999].

Designing a neural network is often trial and error and while there are motivations for some design decisions, their impact on performance is often not clear but needs to be

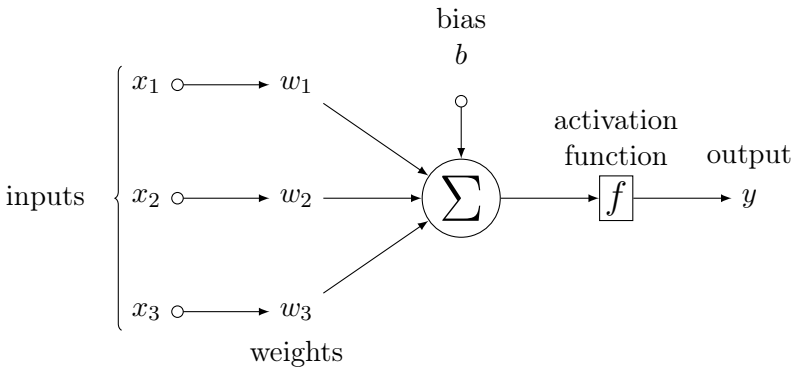


Figure 4.2.: Artificial neuron: smallest unit of a neural network. The weights associated with the neuron and are multiplied with the 3 input features x_1, x_2, and x_3, summed and offset with a bias term. The activation function f, which can be any linear or non-linear function, transforms this value and passes it on either to the output of the network or the next layer.

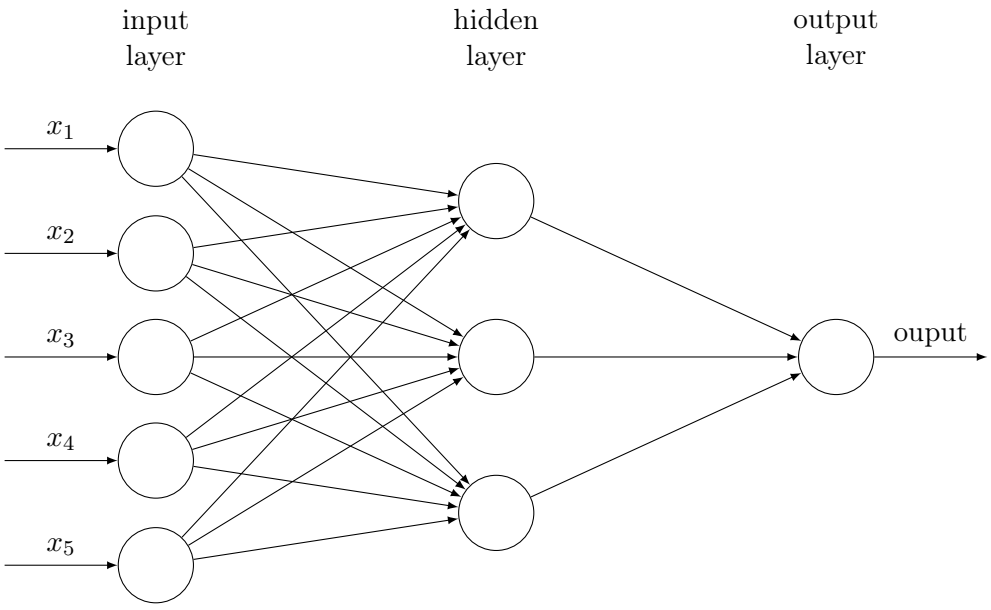


Figure 4.3.: A fully connected neural network consisting of one hidden layer with 3 neurons and an output layer that integrates the information of each unit. If the activation function is a heaviside step function, this neural network can perform classification operations and is called a perceptron.

evaluated empirically such as in [Valle et al., 2017]. Here, I will introduce the building blocks and motivations behind parts of the neural network architectures used for the motion artefact detection (chapter 6).

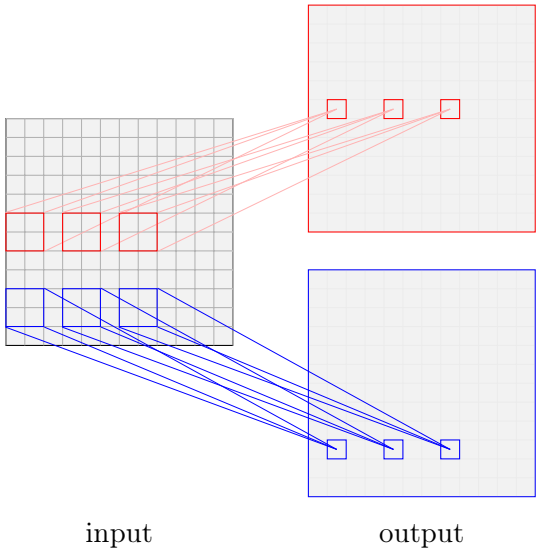


Figure 4.4: Illustration of a convolution layer with two units (red and blue), filter size 2x2 and stride 1. Here, the output layer has the same spatial extent as the input layer, which is typically achieved by zero padding of the input layer.

4.3.1. Convolution layer

Convolutional neural networks for image classification use convolution layers that have typically small rectangular-shaped learnable filters. A filter is applied to a local patch of the input array via element-wise multiplication and summation over all resulting elements (akin to a dot product). The resulting value is added to the bias term and optionally fed into an activation function. This is repeated in a stepwise fashion with a defined step size ("stride") to cover the whole output feature map (see fig. 4.4). For operations with stride one, adjacent output values originate from a patch shifted by one row or one column in the input array. Strides larger than one produce outputs with smaller spatial extent.

In contrast to fully connected layers, where neurons have a connection and associated weight for each input unit (see fig. 4.3), convolution layers reuse the same set of filters by sliding them across the input layer. Hence, filters are learned irrespective to global translation. A convolution layer with Z units produces Z output maps that have a spatial extent determined by the spatial extent of the input layers and by the stride and padding of the convolution layer. If the input to a convolution layer consists of multiple (D) channels, a separate convolution filter is applied to each channel and their output summed to produce a data point in the corresponding output layer. Hence, a convolution layer combines spatially local patterns and correlations across input channels. These operations are repeated for each of the Z output channels using input and output-specific filters. If the filters are of size $M \times N$ then the convolution layer has $M \times N \times D$ parameters per filter, which in total makes $M \times N \times D \times Z$ weights and Z bias terms for the layer. The number of parameters is independent of the size of the input layer and for typical image resolutions much lower than what a fully connected layer would require.

The extracted features can be passed on to another convolution layer that locally combines those features across input channels to create another feature map. The first two layers of a CNNs tend to respond most strongly to local image intensity variations

such as edges and textures while higher levels respond to more complex patterns and show increasing invariance to perspective [Olah, Mordvintsev, Schubert, 2017; Zeiler, Fergus, 2013]. Yet, individual neurons in higher levels do not necessarily encode a single semantic feature to a higher degree than a random combination of neurons, so the information is encoded in the manifold they span [Szegedy et al., 2013].

Alternatives to stacked convolution layers are inception modules that separate the operations across channels from that in the spatial domain [Szegedy et al., 2015; Chollet, 2016]. A very recent alternative are capsule networks [Sabour, Frosst, E Hinton, 2017] that attempt at modelling the "pose" of a feature (translation, rotation and scaling) explicitly but they require different training algorithms.

4.3.2. Pooling layer

Pooling layers locally apply a function to each feature map that reduces a local patch to its average or maximum value. Typically, the maximum activation is used with a filter size of 2x2 and applied with a stride of 2, therefore removing 75% of the parameters (see fig. 4.5). This results in a limited degree of invariance to affine spatial transformations of the input. Most common CNN architectures reduce the width of the network gradually by repeated max pooling filters interleaved with convolution filters. However, pooling layers have been criticised for being a crutch as they could be replaced by convolution filters with stride larger than 1 [Springenberg et al., 2014] or better by explicitly modelling the location and orientation of the features [Sabour, Frosst, E Hinton, 2017] in analogy to the ventral stream in the brain [Poggio, 2011].

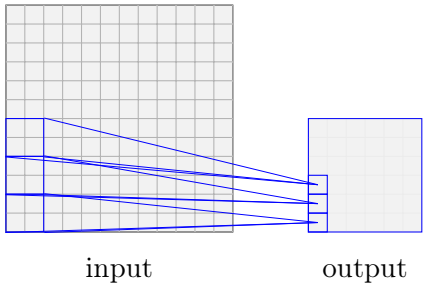


Figure 4.5: Illustration of a pooling layer with filter size 2x2 and stride 2.

4.3.3. Deep, wide, balanced?

The concept of the depth of a network originates from complexity theory and is the largest number of computational steps (layers) from the input to the output [Bengio, 2011]. Theoretically, feed forward neural networks with a single hidden layer and sigmoidal activation functions are capable of approximating any continuous function operating on bounded input to any precision, limited by the number of neurons or the width of the layer [Cybenko, 1989]. This has also been shown for multi-layer networks with any bounded, continuous and nonconstant activation function [Hornik, Stinchcombe, White, 1989; Hornik, 1991].

Deep neural networks are able to compute more complex functions compared to shallow neural networks with the same number of parameters by reusing computations in deeper levels of the network [Montúfar et al., 2014; Pascanu, Montufar, Bengio, 2013]. Note that the ability to compute a function efficiently does not necessarily mean that the model can learn it efficiently [Song et al., 2017]. Although, shallow neural networks can mimic deep neural networks by learning their feature representation, it is difficult to train them from scratch on the original training data [Ba, Caruana, 2013].

A general recommendation by the authors of the Inception architecture in 2015 was to balance width and depth in a network [Szegedy et al., 2015]. Yet, developments since then facilitated very deep [He et al., 2015a] and wide [Zagoruyko, Komodakis, 2016] networks, which outperform previous architectures in large-scale image classification.

Training very deep neural networks has been greatly facilitated by adding "residual" connections between blocks of layers, which help counteract vanishing gradients during training. The performance of those networks is correlated with their depth up to thousands of layers, yet the rate of improvement drops exponentially with the number of layers [Zagoruyko, Komodakis, 2016].

> "Research into convolutional network architectures proceeds so rapidly that a new best architecture for a given benchmark is announced every few weeks to months, rendering it impractical to describe in print the best architecture." [Goodfellow, Bengio, Courville, 2016, chapter 9]

In general the developments in computer vision are too rapid and groud breaking to be summarised meaningfully. When applying a CNN to a new task, it is presumably best practice to use a network design that works well on related datasets.

4.4. Training a neural network and generalisation

Gradient descent allows optimising a classification algorithm by minimising the error on the training data. Similar to the example of logistic regression, neural networks parameters are found via gradient descent [Paola, Schowengerdt, 1995]. CNNs are optimised by an automatic differentiation technique called backpropagation [Rumelhart, Hinton, Williams, 1986; Rumelhart, Hinton, Williams, 1985] or backwards propagation of errors. The training sample is fed to the network to obtain the output value, which is compared to the true label using the cost function (for instance cross-entropy). The true label is propagated sequentially from the output backwards to the input layer to calculate the difference between actual and desired output of each unit in the network. This difference, multiplied by the magnitude of the input to the unit (chain rule), is a linear approximation to the cost function gradient and is scaled by the (negative) learning rate to update all weights and bias terms of the network.

A classification system that perfectly fits the training data would be useless if it could not be applied to unseen data. In other words, machine learning requires not just the optimisation of a cost function on a specific set of data but the ability to generalise [Bengio, 2012].

Learning theory shows that the upper bound of the difference between training and test set error increases with the number of parameters and decreases with the size of the training set (see [Goodfellow, Bengio, Courville, 2016, chapter 5]). However, good vision CNNs often have enough parameters to completely memorise the training data and, despite regularisation, can easily fit random noise [Zhang et al., 2016]. Furthermore, the number of "effective" parameters in neural networks is difficult to estimate [Zhang et al., 2016]. Neural networks have complex cost functions [Nguyen, Hein, 2017] and, theoretically, it should be very hard to train even a 3 node neural network [Blum, Rivest, 1989]. Learning theory can not yet offer model-independent insights into why CNNs generalise well in practice [Zhang et al., 2016; Dinh et al., 2017].

4.4.1. Regularisation

Regularisation are methods that are used to improve the generalisation performance of an algorithm [Kukačka, Golkov, Cremers, 2017]. In optimisation this is often achieved by limiting the solution space in size or by constraining it to certain properties with the aim to generalise better, often at the cost of an increased training error.

In deep learning, regularisation barely limits the networks from perfectly fitting the training set [Zhang et al., 2016] and, empirically, networks perform well even when extremely overparametrised [Poggio et al., 2018] and in the absence of implicit regularisation [Zhang et al., 2016]. However, some amount of regularisation is typically applied and can improve generalisation performance by a few percent accuracy [Zhang et al., 2016]. See [Kukačka, Golkov, Cremers, 2017] for a "taxonomy" of regularisation techniques used for neural networks.

One approach to improve generalisation is early stopping of the training process, which prevents overfitting for some applications with convex cost functions and can be used for

neural networks [Martin, Mahoney, 2017]. However, it is difficult to determine when to stop training and it does not have an effect on all networks [Zhang et al., 2016].

In some architectures, training beyond a decrease in training loss can increase generalisation performance by improving the network's feature representation from an information theory point of view [Shwartz-Ziv, Tishby, 2017]. Shwartz-Ziv, Tishby show that certain neural networks learn in two phases. First, the mutual information between the labels and the network outputs increases, indicating the learning to represent the data. This learning phase is followed by a phase of decreasing mutual information between the input and the network activations. In the latter stage, networks seem to forget learned but unimportant connections in the training data, leading to a better generalisation performance. However, these findings might be network architecture specific [Amjad, Geiger, 2018].

The choice of loss function, learning rate, batch size, dataset, and network architecture itself influence the cost function landscape and how it is traversed and therefore implicitly have an impact on the generalisation performance [Jastrzębski et al., 2017]. For instance, convolution layers use their filters globally, which constrains the network to learn filters that are invariant to translation.

Stochastic or mini-batch gradient descent updates the parameters of the network based on little data and results in noisy trajectories. However, traversing the cost landscape using all training data (batch gradient descent) does not necessarily improve the final accuracy but requires much smaller learning rates on non-linear cost function landscapes [Wilson, Martinez, 2003]. Often, networks trained with very small batch sizes generalise better, hence noisy steps can be seen as regularisation. This gain, however, is offset by an increased computational cost of updating the network parameters for each sample independently [Goodfellow, Bengio, Courville, 2016, chapter 8].

Batch normalisation layers were designed to stabilise training. They normalise their input, the activations of the previous layer for all samples of the training batch, to zero mean and unit variance via a batch-specific affine transformation (scale and offset). These normalised activations are then transformed with an affine transformation that is independent of the training batch but are learned parameters of the layer. This batch-wise normalisation and re-scaling stabilises learning by factoring out fluctuations in the magnitude of the activations when weights of other layers are changed and leads to faster training [Ioffe, Szegedy, 2015]. Although they were not intended for regularisation, batch normalisation layers improve generalisation performance in many cases [Zhang et al., 2016; Ioffe, Szegedy, 2015] and are part of high-performing network architectures [Szegedy et al., 2015; He et al., 2015a].

Dropout regularisation [Srivastava et al., 2014] is a simple and popular technique to prevent overfitting on small and noisy datasets [Jindal, Nokleby, Chen, 2017]. During training, it randomly sets a fraction of weights in a network to zero to avoid units to become dependent on specific combinations of features thus favouring features that are robust to small changes in context.

Another regularisation technique is forcing the network to learn classification rules that do not perfectly separate the training data but leave some degree of uncertainty [Szegedy et al., 2015]. This can be achieved with label smoothing or by penalising low

entropy classification distributions [Pereyra et al., 2017].

Regularisation techniques can also increase the effective size and variance of the training data by transforming the data directly in ways that preserve the characteristics of the data (data augmentation) or by transforming the extracted representation in the network (feature augmentation). If data transformations can be applied, they often outperform feature augmentation methods [Wong et al., 2016]. Common image augmentation strategies are erasing [Zhong et al., 2017], cropping [Krizhevsky, Sutskever, Hinton, 2012], flipping [Simonyan, Zisserman, 2014], padding and rigid or non-linear transformations of the input images. See [Kukačka, Golkov, Cremers, 2017] for an overview of data-related regularisation techniques.

4.4.2. Learning from imbalanced data

Compared to balanced datasets of similar size, imbalanced data or data with skewed class distributions can introduce biases and reduce generalisation performance of classification algorithms. Although in practice not always the case [He, Garcia, 2009], it is intuitively clear that if a minority class with a complex feature space is represented by very few samples in the training data, there might not be sufficient patterns present for an algorithm to learn a generalisable representation of that class.

This overfitting of the minority class can be especially problematic if noise is present in the data. Furthermore the difference in data densities between classes can cause a classifier to learn a decision boundary that is biased as the cost of misclassifying the minority class is proportionally smaller. This gets exacerbated in the presence of small disjuncts or overlapping class boundaries and can be problematic if there is a mismatch between distributions in training and test data (sample selection bias or dataset shift) [López et al., 2013].

It is instructive to simulate this on a one or two dimensional dataset and plot the decision boundary as a function of between class imbalance. As in fig. 4.1, the generating distributions have slightly overlapping class boundaries but equal class densities. Samples with varying degrees of class imbalance are drawn from these distributions and 3 classifiers are fit as above: a linear logistic regression model and two neural networks with one hidden layer containing 3 and 5 neurons, respectively.

On the balanced sample, the neural network can separate the classes with a nonlinear boundary. The logistic regression finds a decision boundary that best separates the data with a straight line, which is a crude approximation of the half-moon shaped boundary between the two classes but the best separation a linear classifier can achieve. However, this boundary shifts towards the minority class with increasing between-class imbalance and would achieve low performance if tested on a non-skewed dataset (see fig. 4.6). With increasing class imbalance, the neural network becomes more susceptible to noisy samples of the majority class and learns a class boundary that has little in common with the original sample densities.

The last row shows the classifier training and validation performance (accuracy and area under the ROC: AUROC) as a function of the majority fraction with 95% bootstrap confidence intervals for 10 independent test datasets. For each bootstrap estimate, 1000

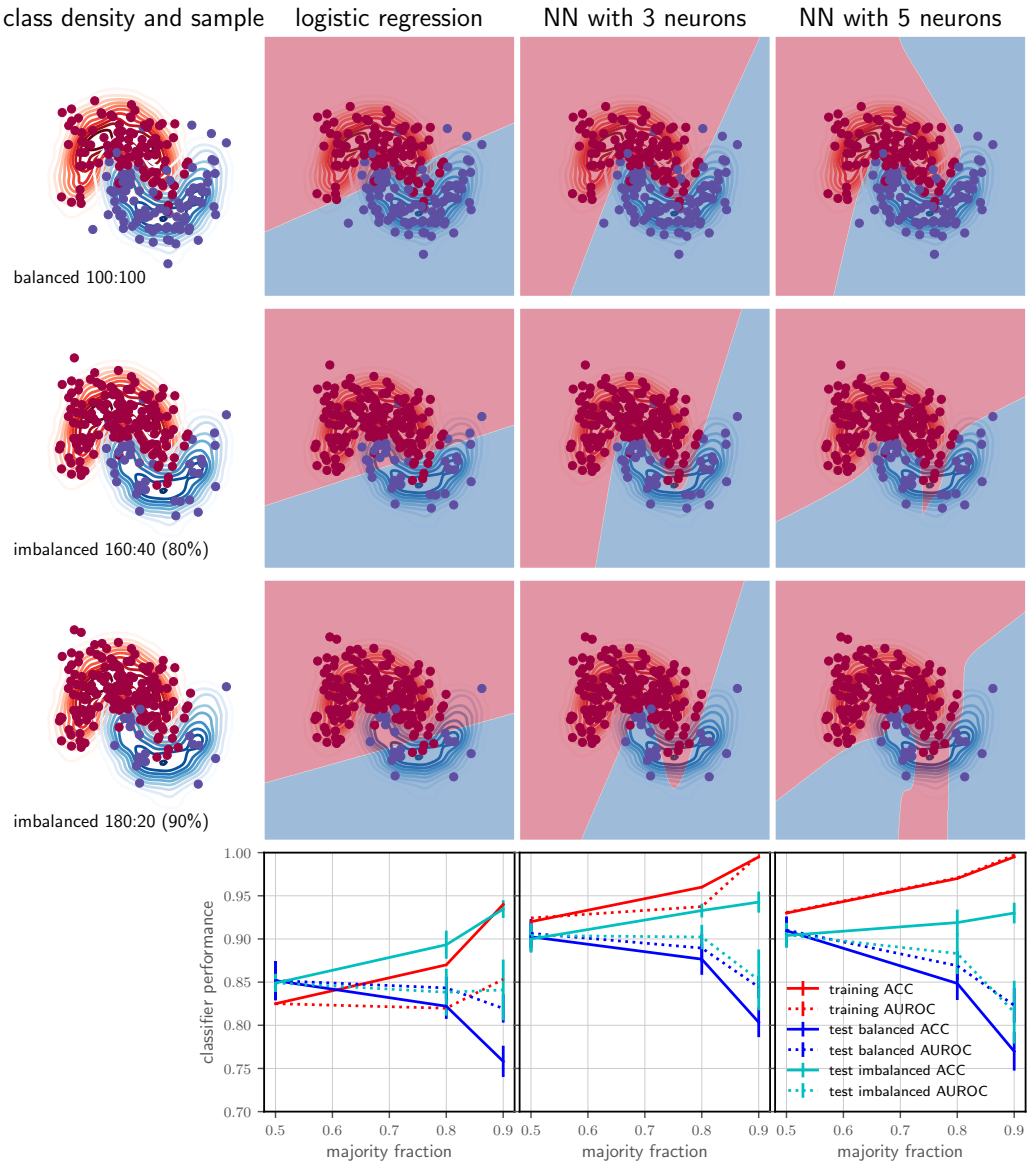


Figure 4.6.: Classification performance in the presence of a skewed sample distribution: the top row shows (as in fig. 4.1) two sampling probability distributions in the shape of half-moons, from which samples are drawn for the blue and red class. The background colour in the images on the right indicates the learned separation of the parameter space for a logistic regression and two neural network classifiers. The training sample is skewed for row 2 and 3 with 80% and 90% of the samples belonging to the red class. The decision boundary is increasingly biased for higher sample imbalance. The 5 neuron neural network degrades faster than the 3 neuron network indicating its tendency to overfit.

samples are drawn from the original distribution, either distributed equally across classes

("test balanced") or with the same imbalance as the training sample ("test imbalanced"). Test accuracy on a similarly imbalanced dataset increases with increasing imbalance but this is owed to accuracy being a poor measure for imbalanced datasets. AUROC decreases for all classifiers with increasing imbalance, irrespective whether the test data is balanced or imbalanced.

Many techniques have been developed to improve learning from imbalanced data [He, Garcia, 2009], most notably sampling techniques and cost-sensitive learning [López et al., 2012], kernel-based and active learning techniques, and ensemble methods [Galar et al., 2012; Khoshgoftaar, Van Hulse, Napolitano, 2011]. In the following I will describe some ideas that are most relevant with regard to CNNs as used in chapter 6, and point the reader to the reviews [He, Garcia, 2009; López et al., 2013; Japkowicz, 2000] for more details.

King et al. proposes to balance the classes by weighting their contribution to the cost function [King et al., 2001]. Assuming an imbalanced dataset of N samples with N_i samples in class i and $i = 1...M$, one can weight the contribution of samples from class i with $w_i = N/(M * N_i)$, satisfying the conditions that the effective number of data points is equal across classes ($w_1 * N1 = w_2 * N2 = ... = w_M * N_M$) and that the total number of effective data points remains the same as in the non-weighted case ($w_1 * N_1 + ...w_M * N_M = N$). A similar approach is taken for cost-sensitive learning of neural networks in [Kukar, Kononenko, 1998], where the step size of the gradient descent is weighted by the misclassification cost.

In the case of logistic regression, binary trees and some ensemble methods, it is possible to tweak decision thresholds to adjust to class imbalance after training ("threshold-moving") [Collell, Prelec, Patil, 2016]. This has been applied to neural networks [Saerens, Latinne, Decaestecker, 2002] but can not account for learned distorted non-linear class boundaries as demonstrated by the decrease in AUROC in fig. 4.6.

Alternative approaches use oversampling of the minority class, undersampling of the majority class or hybrids. However, undersampling looses information and repeated sampling of the same data creates artificial high density areas in the feature space, which might cause a non-linear classifier to overfit to the samples of the minority class. Techniques such as SMOTE [Chawla et al., 2002] and ADASYN [He et al., 2008] address this issue by creating synthetic samples by combining existing samples that are in areas where the classifier might most benefit from a higher sample density. These approaches require the ability to create synthetic samples or, in the case of multi-layer networks, at least synthetic features generated from intermediate layers.

4.4.3. Transfer learning

Training a neural network from scratch with high generalisation performance typically requires tens to thousands of examples to map the labels to the training data and to generalise well to unseen data [Vinyals et al., 2016; Hochreiter, Younger, Conwell, 2001]. Humans on the other hand seem to learn richer representations than machine learning algorithms from only one or very few examples [Lake, Salakhutdinov, Tenenbaum, 2015].

Given large amounts of training data, deep neural networks learn to extract useful

characteristics or sub-structures of the training data and this ability can be transferred to a different but related task domain if these extracted representations generalise to the new task [Bengio, 2011; Andrychowicz et al., 2016]. Transfer learning allows transferring the learned feature extraction to related domains where data is limited, hard to collect [Pan, Yang, 2010] or even completely absent [Bengio, 2011]. In image-classification tasks, typically only the final layers of the network that map the features have to be learned from scratch to the new target function. However, if enough training data is available, fine-tuning of more or all layers often increases classification performance [Lamblin, Bengio, 2010].

Interestingly, many medical image classification applications of transfer learning have successfully used models trained on large general purpose image datasets and domain specific pre-training is not necessarily better or can lead to worse performance than networks pre-trained on large general purpose datasets [Menegola et al., 2016]. Ruder, Plank have used statistical methods to select data suitable for transfer learning and emphasize that not only domain similarity but also diversity in the data are crucial for successful transfer learning [Ruder, Plank, 2017].

Chapter 5

Binary classifier performance evaluation on imbalanced data

Contents

5.1. Introduction

Classification algorithms play an increasing role in processing and analysing medical data. In chapter 6, we develop a neural network classification algorithm to detect motion corrupted diffusion data with the aim of preventing severe motion artefacts from affecting subsequent analysis. A fundamental question in building and comparing these algorithms is how to compare their performance if the cost of misclassification is not known or if it depends on the target application. In neonatal imaging, for instance, a certain degree of redundancy is incorporated in the sequence design to account for motion. Applying the same classifier to a dataset with less redundancy, e.g. due to scan

abortion, potentially changes the cost of misclassifying an acceptable volume as unusable. Furthermore, medical data collection is commonly focused on specific groups (patients or healthy controls), expensive, and often invasive, resulting in small and imbalanced datasets. Given a specific test set, assessing and comparing the performance of classification algorithms is surprisingly non-trivial when the cost of misclassification is not known [Adams, Hand, 1999; Parker, 2013; Japkowicz, Shah, 2011] and in the presence of skewed class distributions [López, Fernández, Herrera, 2014]. Classification performance analysis research is spread across disciplines that use different jargons [Lavesson, Davidsson, 2007] and lacks consensus [Jamain, Hand, 2008], especially with regard to interpretability and significance of differences [Japkowicz, Shah, 2011; Steyerberg et al., 2010].

In practice, classifier performance is dependent on the chosen metrics [Weiss, Provost, 2003] and it is common to report multiple performance metrics. However, any single measure of performance is limited, might oversimplify [Drummond, Japkowicz, 2010; He, Garcia, 2009], and makes different implicit or explicit assumptions about the cost of misclassifying samples of each category [Parker, 2013; Hand, 2009]. Unfortunately, model selection and comparison is often based on metrics that make poor assumptions about the problem [Adams, Hand, 2000] or use methods that are biased [Forman, Scholz, 2010] or yield highly variable rankings [Efron, Tibshirani, 1997].

This chapter serves as a brief introduction to the most commonly reported performance metrics (section 5.2) and demonstrates how these metrics' sensitivity to classifier performance changes for skewed data distributions (section 5.3). The simulations presented here aim to give an intuition about performances reported in chapter 6 and how they depend on the class imbalance naturally present in motion corrupted data and in the uneven distribution of diffusion weightings in any dataset. Finally, I give a brief overview of non-parametric generalisation estimation techniques (cross-validation and bootstrap) and how they are used in deep learning.

5.2. Background: Binary classifier performance metrics

5.2.1. Point measures

	actual positive	actual negative
predicted positive	true positive (TP)	false positive (FP)
predicted negative	false negative (FN)	true negative (TN)

Table 5.1.: Confusion matrix. T and F stand for true and false, indicating the agreement between classifier and ground truth labels for samples that the classifier identified as positive (P) or negative (N). Hence, the P and N in FP and FN do not refer to the true class labels but to the classifier's prediction.

Accuracy measures the closeness of the prediction to the ground truth labels, irrespective of the class label and distribution. Expressed in units of the classification confusion matrix (see table 5.1), accuracy is the sum of true positives and the true negative predictions normalised by the number of samples $(TP + TN)/(TP + TN + FP + FN)$.

Precision is the fraction of correctly labelled positive labels of all positive classifications $TP/(TP+FP)$ and is a measure of the exactness of a classifier in retrieving positive samples. Recall or sensitivity, is the proportion of correctly labelled positive samples relative to the total number of positive samples $TP/(TP+FN)$. It measures the completeness in the context of retrieving all positive samples.

The F_β score or F-measure is the weighted harmonic mean of precision and recall:

$$F_\beta = \frac{(1+\beta^2)\cdot \mathrm{PREC}\cdot \mathrm{REC}}{(\beta^2\cdot \mathrm{PREC}) + \mathrm{REC}} \tag{5.1}$$

F_β is 1 only for perfect precision and recall and goes to zero if either precision or recall go to zero. For $\beta < 1$, precision is weighted higher than recall.

Given that the goal is to estimate the performance of a binary classifier, accuracy might be an obvious choice to measure the agreement between the test set and the classification result [Ling, Huang, Zhang, 2003]. However, accuracy assumes an equal cost of misclassifying positive and negative samples and is sensitive to the ratio of positive and negative samples, which makes it flawed for comparisons across datasets [Demšar, 2006]. On an imbalanced dataset, such as the motion artefact dataset in chapter 6, accuracy is a poor choice to assess the quality of a classification algorithm, as a classifier that always predicts the majority label can achieve high accuracy but is practically useless. Parker does "not even bother to discuss it" in his performance metric comparison [Parker, 2013].

Precision and recall do not take the true negative values into account and therefore, neither does the F-measure. This can be irrelevant to measuring information retrieval systems that report the k most relevant samples of a large pool of samples but is not ideal for binary classification [Powers, 2011]. Alternative measures based on the confusion matrix are the Cohen kappa (Cohen_κ) and Matthews correlation coefficient. The Cohen kappa score measures the agreement between two raters, corrected for agreement expected if both annotations were random label vectors. However, Cohen_κ scores are hard or impossible to interpret and recently, authors of Cohen_κ variants discouraged the use of all Cohen_κ measures [Pontius Jr, Millones, 2011]. Here, I report the scores computed by scikit-learn [Pedregosa et al., 2011] for completeness and comparison reasons.

The Matthews correlation coefficient (MCC) [Matthews, 1975] is an alternative to the F-measure and Cohen_κ score. The MCC is equivalent to the Pearson correlation coefficient between two Bernoulli random variables and can be calculated using all entries in the confusion table.

$$\mathrm{MCC} = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5.2}$$

In contrast to the other measures discussed here, the MCC is a correlation coefficient and ranges from -1 to 1. Values below 0 indicate disagreement beyond the level of chance. The MCC is an often recommended measure for reporting performance on imbalanced data if the confusion matrix has to be represented as a single number [Powers, 2011; Shi et al., 2010; Chicco, 2017; Boughorbel, Jarray, El-Anbari, 2017]. However, Powers suggests using the MCC not for performance comparisons between classifiers but as a

measure of changing "behaviour" of a classifier when tested on a differently skewed test set [Powers, 2012].

Accuracy, F_β, MCC and Cohen$_\kappa$ scores are set-based measures that do not take the rank of the classification results with respect to the ground truth labels into account. When they are used for assessing continuous labels produced by a probabilistic classifier, then the class boundary needs to be defined using a threshold t on the classification vector. However, in practice, the optimal threshold is application dependent and thresholding looses discriminative information that a "proper scoring rule" [Gneiting, Raftery, 2007] can use. Even if a best guess about the misclassification cost of each class and therefore optimal threshold can be made, it is more plausible to define the cost using a distribution instead of a single point [Hand, 2009].

5.2.2. Binary integrated measures

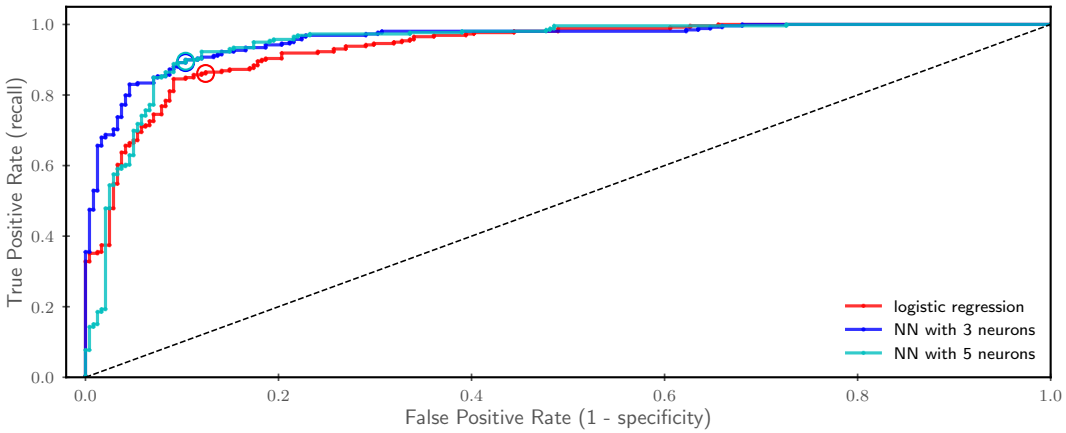


Figure 5.1.: Receiver operator curves of the classifiers shown in fig. 4.1 evaluated using 500 test samples. At the chosen operating point (probability threshold 0.5, circles), the neural networks perform nearly equally well, outperforming the logistic regression. However, in the high specificity (low false positive) area of the plot, the logistic regression and the neural network with 3 units have higher recall than the neural network with 5 units. Hence the ranking depends on the threshold, which is determined by what an acceptable false positive rate is. The area under these curves is a performance measure that is independent of the misclassification cost.The dashed line indicates performance equal to random chance.

The area under the receiver operator curve (AUROC), average precision (AP), and the H-measure are integral measures and use information about the classification label ranking instead of applying a single classification label threshold. The receiver operator curve can be generated by plotting recall versus the false positive rate (1 - specificity) for each class threshold between 0 and 1 (see fig. 5.1). The AUROC is the area under this curve and represents the probability that a classifier assigns a higher value to a sample of the positive class than to a sample of the negative class [Mason, Graham, 2002]. In scikit-

learn, which is used throughout this thesis, the AUROC is calculated via the trapezoidal method using all unique classification values in the classification vector as thresholds.

For classification algorithms that produce probabilistic labels, accuracy is empirically [Bradley, 1997] and provably [Ling, Huang, Zhang, 2003] less discriminating than the area under the receiver operator curve. Probabilistic algorithms trained with AUROC as the loss function achieve higher AUROC and higher accuracy than when trained using accuracy [Ling, Zhang, 2002].

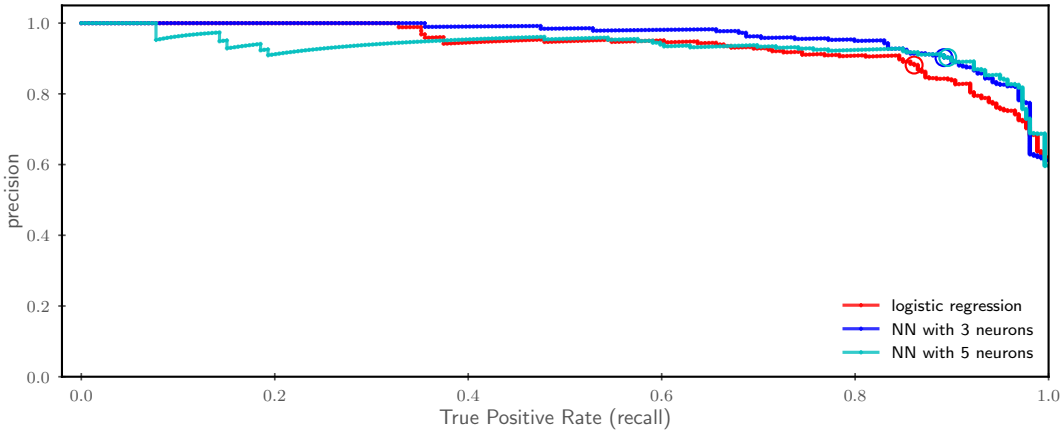


Figure 5.2.: Precision recall curves for the same classification results as shown in fig. 5.1.

Average precision (AP) is defined as the area under the precision-recall curve and is typically used to compare information retrieval algorithms that return a fixed number of samples that most likely represent samples of the positive class. Given a vector of probabilistic classification labels, one can sort it in descending order and calculate the precision and recall for every value in the vector. For a fixed length sample vector, AP can be interpreted as the average precision measured whenever a positive sample is observed. Similar to the receiver operator curve, the precision recall curve (fig. 5.2) can be generated by varying the class-boundary of the predicted labels from 0 to 1. However, there exist at least 8 different algorithms to compute AP [Boyd, Eng, Page, 2013]. Implementations differ in how they treat coordinates with different precision but equal recall. I use the scikit-learn implementation, which uses a step function integral, summing the precision values over all thresholds, weighted by the increase in recall from the previous threshold

$$\mathrm{AP} = \sum_{t_i} (\mathrm{REC}_{t_i} - \mathrm{REC}_{t_{i-1}}) \mathrm{PREC}_{t_i} \tag{5.3}$$

The AUROC has been criticised for being "incoherent" by implicitly weighting misclassifications of the positive and negative class differently, depending on the classifier itself [Hand, 2009; Parker, 2013]. Hand proposed an alternative cost function, the H-measure. I will briefly summarise the motivation behind the H-measure, for the derivation see [Hand, 2009] and for a discussion see [Parker, 2013]. The loss function $L(t)$ at a given

threshold depends on the distributions of scores y that the classifier produces for each class: $F_-(t) = p(y >= t|-)$ and $F_+(t) = p(y < t|+)$ and on the fractions of samples in the classes π_- and π_+

$$F(t) = \pi_- F_-(t) + \pi_+ F_+(t)$$

with $\pi_- \in [0, 1]$, $\pi_+ \in [0, 1]$ and $1 = \pi_- + \pi_+$.

This assumes an equal weighting between misclassifying samples of either class. Hand defines a relative "cost" of misclassifying a sample of the negative class $c \in [0, 1]$; $1 - c$ is the cost of misclassification of the opposite class.

$$F(c, t) = c\pi_- F_-(t) + (1 - c)\pi_+ F_+(t) \tag{5.4}$$

For any given cost c, the loss can be minimised with a specific threshold $t_c = \text{argmin}_t(F(c, t))$. Instead of using a single cost, Hand argues to integrate eq. (5.4) over a range of cost values weighted by a function $w(c)$ that expresses domain knowledge about the expected distribution of cost values. Using this cost function formulation, Hand shows that the AUROC implicitly uses a classifier-dependent cost distribution $w(c)$. However, a classifier-dependent misclassification cost distribution does not make sense, as it should be application-specific and fixed.[1] The H-measure on the other hand uses an explicit distribution of misclassification weightings for each class.

To make results comparable, the original method [Hand, 2009] proposed to use a symmetric distribution (Beta(2, 2)) as a universal weighting function. However, the default H-measure was later refined to take the prevalence of positive and negative labels into account (Beta($\pi_+ + 1$,$\pi_- + 1$)) [Hand, Anagnostopoulos, 2014]. The Kolmogorov Smirnov statistic has a symmetric cost of misclassifying all cases of one class but being completely correct on the other class. The refined H-measure is an extension of this statistic, which would use a cost value $c = \pi_+$, to a distribution with mean π_+. Here, the H measure that uses the symmetric beta distribution is denoted as H_1, the class-imbalance sensitive version as H.

Note that the conclusion that the AUROC is incoherent has been challenged in [Ferri, Hernández-Orallo, Flach, 2011] who use a different definition of loss. Ferri, Hernández-Orallo, Flach come to the conclusion that AUROC assumes a uniform distribution of misclassification costs and a uniform distribution of thresholds. See [Parker, 2013] for a brief discussion of both methods.

Example To illustrate the behaviour of different measures, consider the true class labels $\mathbf{y}_{\text{true}} = (0, 0, 0, 0, 1, 1)$. A classifier that produces the labels $\mathbf{y}_1 = (0, 0, 0, 0, 0, 1)$ or $\mathbf{y}_2 = (0, 0, 0, 1, 1, 1)$ has the same accuracy (0.83) but average precisions 0.7 and 0.67, respectively. The AUROC and H-score assign lower scores to the classifiers that misclassify the minority class: AUROC ($\mathbf{y}_1$) = 0.75, AUROC ($\mathbf{y}_2$) = 0.88, H($\mathbf{y}_1$) = 0.43, H($\mathbf{y}_2$) = 0.5. The prediction $\mathbf{y}_3 = (0, 0, 0, 0.1, 0.2, 1)$ has an AUROC, AP and H-measure

[1] Akin to comparing two peoples' height in units of the length of their feet: the metric would depend on the object measured.

score of 1 as it perfectly preserves the true label rank but an accuracy of 0.67 if thresholded at 0.5. The MCC values for y_1 and y_3 are 0.63, that for y_2 is 0.71.

5.3. Simulations: performance estimation on imbalanced data

5.3.1. Introduction

The performance of a classifier for a specific application depends on the class balance and on the cost of misclassifying either category. For instance, assume that the positive class represents the presence of a disease, for which the treatment is relatively safe and cheap but not treating it would be dangerous and costly. Let p_+ be the likelihood of yielding a correct result when the sample is of the positive class and p_- be the corresponding likelihood for the negative class. In this scenario, a free and harmless medical test that has a high probability of success of correctly identifying diseased subjects (p_+) would be preferential to one with lower p_+, nearly irrespective of their respective chance for false positives. If the cost of the test (or that of the treatment) increases, p_- becomes more important, especially if the population consists of many negative samples.

By design, a single number can not represent the full information about the classification performance on the majority and minority class and give information about the class distribution. While most metrics have an intuitive interpretation, it is difficult to compare two classifier's performance as the metric value and its sensitivity to changes in performance on detecting positive or negative samples can depend on the test set imbalance.

To put the classifier performance values into perspective, I simulate a binary "classifier" with well-defined performance on the positive and negative class and report classification performance metrics using 7 of the most common performance metrics on balanced and unbalanced data: accuracy, F_β-score ($\beta = 1$ and 0.1), Matthews correlation coefficient (MCC), Cohen kappa (Cohen_κ), area under the receiver operator curve (AUROC), average precision (AP), and the H (and H_1) measure. This allows comparing performance metrics as a function of classifier performance and imbalance while controlling for the classifier's ground-truth performance.

5.3.2. Simulations

Different binary classification performance metrics implicitly or explicitly weigh the cost of misclassifying samples of the two classes differently and their score can depend on the prevalence of samples from the positive π_+ and the negative class π_-. Therefore, comparing algorithms across different datasets using dataset-dependent performance measures, requires a calibration of metric values.

To simulate classifiers with a class-specific performance, I use a Bernoulli trial for each class with a class-dependent probability of success (p_+ and p_-). Each simulated classifier returns a total of 100'000 draws from those class-specific binomial distributions $\mathbf{y} = (\mathcal{B}(1, p_-), \mathcal{B}(1, p_-), ..., \mathcal{B}(1, p_+), \mathcal{B}(1, p_+))^T$. p_- and p_+ are limited to $[0.5, 1.0]$, as

labels of classifiers performing worse than chance can simply be inverted. Without loss of generality, I will assume 0 as the labels of the negative and 1 for the positive class. This setup allows simulating class imbalance with a well defined classifier performance and sample size by varying the fraction of data in one of the categories. The maximum simulated imbalance is 1% minority class size (1000 samples) and was decreased to 10% and equally sized groups.

To motivate this simulation, consider an image classification task of categorising two types of trees. Assume a given classifier to perform poorly if the images contain only the trunk and branches but performs well if they contain leaves. p_+ and p_- are different if one type of tree is foliated for a shorter period than the other type or in other words, given a selection of images taken at random time points, it is less likely to correctly classify the former type of tree.

5.3.2.1. Comparing performance values

Figure 5.3 shows performance metric values for classifiers with varying probability of success in each category. Each 2D plot spans a grid with varying p_+ and p_- and shows the associated performance metric values. Different class imbalance scenarios are organised in columns.

Contour lines in fig. 5.3 illustrate which p_+ and p_- combinations are equivalent for any given metric. The orientation of those lines highlights the varying assumptions about misclassification costs in each category: metrics with horizontal or vertical contour lines are insensitive to one of the classes.

With the exception of AUROC values, all investigated performance measures values are dependent on the class balance, π_+, and π_-. The other metrics' sensitivity to changes in the performance of classifying samples of the minority and the majority class changes with the class imbalance. Hence, if the class imbalance is unknown, AUROC provides values that are directly comparable. For $\pi_+ < \pi_-$, the H-measure and the AUROC have similar gradients with respect to the class-specific classification performance. In contrast, the H_1, Cohen$_\kappa$ and the MCC show a class performance sensitivity similar to accuracy when the classifier performs well on the majority class. They are dominated by the majority class when high performing classifiers are compared on imbalanced data.

A class skew dependence makes comparison across datasets challenging. Across different class-imbalance scenarios, the change of a performance metric's gradient with respect to p_+ and p_- leads to possibly different decisions about classifier rankings. This is acceptable (and desired) when the cost of misclassification is known. However all metrics except of the AUROC make assumptions about this cost and these assumptions change with changing class imbalance.

5.3.2.2. Uncertainty due to noisy test data

Due to the stochastic nature of the classifier simulations, the predicted labels (and the predicted class distribution) varies between draws. These probabilistic fluctuations are inherent to the Bernoulli trial and would not occur in a classifier that consists of only

deterministic elements. However, evaluations in chapter 6 use random distortions of the test images (test set augmentation) that slightly affect the predicted label; but the ground truth label and the classifier remain fixed. Note that this is different from random fluctuations in the test set (data and labels) due to, for instance, bootstrap sampling from a test data pool as ground truth and predicted labels do not commute for all performance measures ($\mathrm{AUROC}_{\hat{y}}(y) \neq \mathrm{AUROC}_{y}(\hat{y})$).

To illustrate this using the example of the tree-classification problem: if the tree photos were captured at the same time but independently for each classifier (for instance using a Polaroid film camera), each classifier would see slightly different contrasts, colours and lighting conditions. Both, classifiers and trees, are the same but the predicted class probability is likely a function of image contrasts and therefore will differ between both classifications, which in turn changes the associated performance values.

This variation is due to changing test conditions and how they affect the classifiers' prediction. If this fluctuation is large compared to the gradient of the metric with respect to classifier performance, then it requires a large number of test evaluations to accurately rank the classifiers with respect to their (average) performance. Furthermore, performance metrics that use nonlinear functions (F_β score, H-measure) can amplify the ranking uncertainty in flat areas of the performance landscape but decrease it in steep areas. For a visual impression of this variability using only one trial, see fluctuations in the contour lines in fig. 5.3.

Repeating the simulations multiple times allows investigating how useful each metric is for ranking classifier performances in the presence of test-noise. Correctly ranking similar-performing classifiers requires more test repetitions if a metric fluctuates substantially between test repetitions but changes little with respect to changes in classifier performance. This is especially relevant for the assessment of algorithm performance on small test sets, common in medical imaging.

Using repeated testing, confidence interval of the performance measure value can be estimated to assess the variability of a given classifier due to test data augmentations (see fig. 5.4). By extrapolation of the confidence intervals between adjacent points in the π_-, π_+ landscape, it is possible to estimate the uncertainty in π_- and π_+ that this variation causes.

Confidence intervals in fig. 5.4 are calculated using t-statistics and the logit-transformation ($\log\left(\frac{x}{1-x}\right)$), which limits confidence intervals to the range [0,1] and has been shown to yield accurate confidence intervals for mean average precision in the setting of information retrieval and bootstrapping [Park, 2011].

Due to the different (implicit or explicit) relative weighting of π_- and π_+ between performance metrics, this uncertainty becomes a function of class imbalance, π_-, π_+, and metric; hence plotting this relation requires more dimensions than can comfortably fit on this page. Therefore, fig. 5.4 shows horizontal and vertical "cross-sections" of the 2D colour maps shown in fig. 5.3. In each column, performance of the classifier is fixed for one of the classes but varied for the other. For any point along the metric performance curves in fig. 5.4, the horizontal extent of the estimated confidence band indicates the uncertainty in either π_- or π_+ due to the test-variability. In other words, the horizontal

extent of the 95% confidence bands in fig. 5.4 indicates the plausible range of p_+ or p_- values that a classifier with a fixed performance could have, when evaluated 9 times; it approximates the range of classifier performance values that can not be distinguished from each other with less than 10 test repetitions.

Accuracy, F_β and Cohen_κ have the highest uncertainty for ranking high-performing classifiers if they differ in their performance on the minority class. For the purpose of this work (a moderate class imbalance and high-performing classifiers), the most reliable metrics for ranking in the presence of test-data augmentation are AUROC, MCC and H-measure, followed by AP and H_1.

Note that the expected variance of the sample mean of n independent draws from a binomial distribution $\mathcal{B}(1, p)$ is $\frac{1}{n}p(1-p)$. For n=100'000 draws, the standard error of the sample mean is at most 0.16% for p=0.5 and below 0.1% for p=0.9. For n=1000 draws, the respective standard errors of the sample mean are 10 times larger. Hence, using two Bernoulli trials of constant total size, the re-test variability is performance and class imbalance-dependent. In future work, these simulations could be repeated with constant re-test variability, to allow comparison across class-imbalance domains, and could be extended with a more fine-grained numerical analysis of the width of the confidence bands or with an analysis of the variability of classifier ranks.

5.3.2.3. Rank-preserving label noise

For assessing a probabilistic binary classifier, a performance metric should be robust to label noise that does not affect the order of the predictions with respect to the ground truth labels. In the example given above, $\mathbf{y}_4 = (0.1, 0, 0, 0, 1, 1)$ should yield equal performance to $\mathbf{y}_5 = (0, 0.1, 0, 0, 1, 1)$ as the data can still be perfectly separated into the two classes with any threshold between 0.1 and 1.

To test the tolerance of the implementations of the performance metrics to this kind of noise, the above simulations were repeated with additive noise in the range [0,0.5) and (-0.5,0] for negative and positive classification labels. Noise was added to the negative and subtracted from the positive class. It was drawn from a beta distribution with shape parameters $a = 1$ and $b = 10$, which has an average value of 0.092 and a cumulative density of 0.999 at 0.5, and values above or equal to 0.5 were set to a random number uniformly sampled between 0 and $0.5 - 10^{-8}$ to not affect the ranking of classification labels with respect to the ground truth labels. The simulations were repeated with a second distribution that is generated in the same way but with noise closer to white noise (beta distribution with $a = 1$ and $b = 4$). Both distributions are shown in fig. 5.5.

H measure and AUROC show no visible difference in re-test variability for classification labels with or without rank preserving noise (not shown). However, average precision yields different values compared to the noiseless binary classifier. AP scores are systematically lower on noisy labels than on binary labels (fig. 5.6B). The bias exceeds 8% of the noise-free AP value in the case $\pi_1 = 1\%$ and depends on the relative performance of the classifier on either class. Hence, when comparing two different classifiers, the worse-performing classifier might score higher if it produces less fine-grained labels. Figure 5.6C shows the relative difference in AP values between simulations with differently

distributed label noise. For these two classifiers, AP shows higher variability compared to the comparison of two simulations without label noise (fig. 5.6A) but has no detectable bias. The implementation is therefore presumably biased in the presence of different granularity in the label vectors.

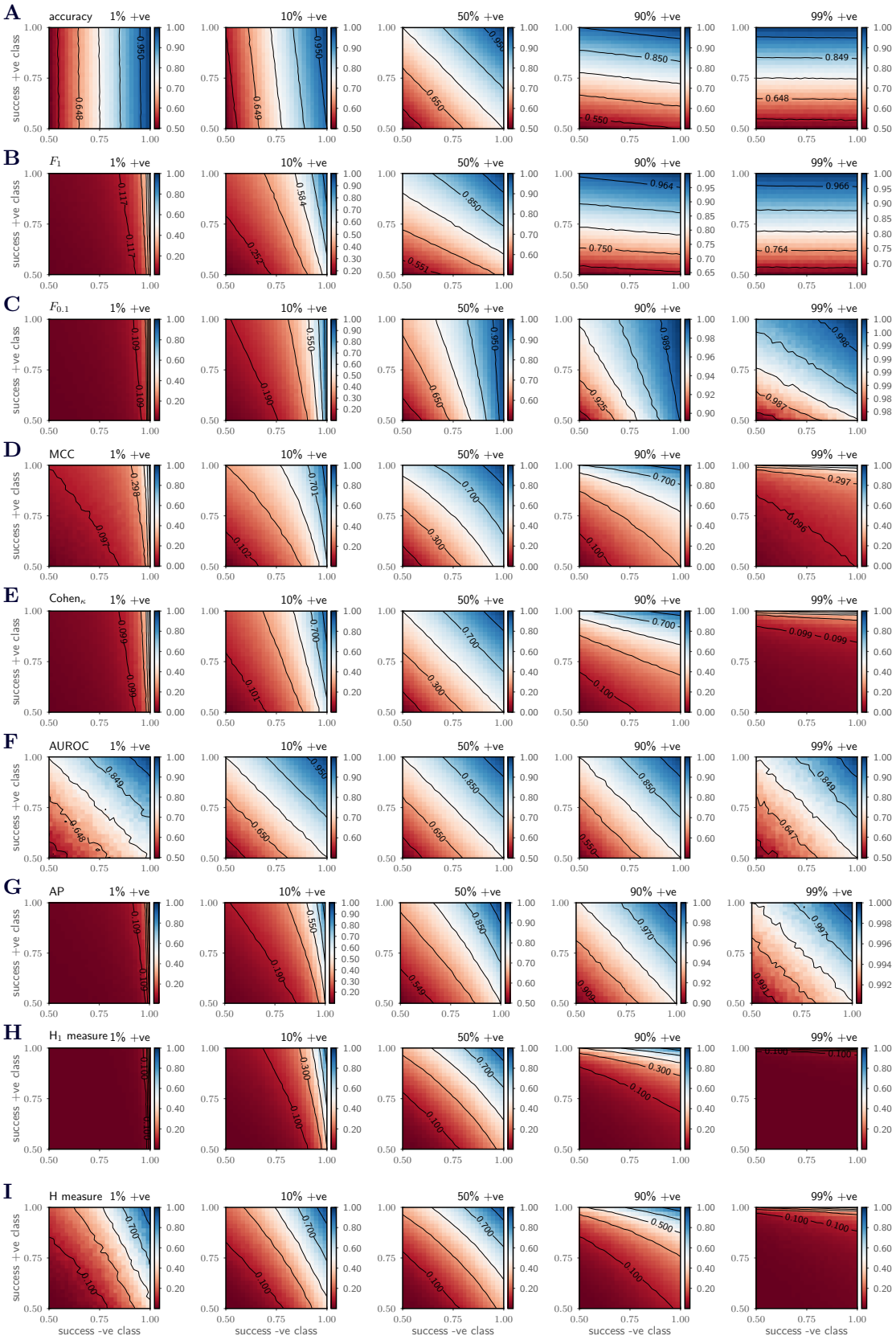


Figure 5.3.: Comparison of different classifier performance measures for simulated classifiers (rows A to I) with varying degrees of class imbalance between the positive (feature = 1) and negative (feature = 0) class (columns). Intensity represents the metric value for a classifier with a specific likelihood of success in the positive and negative class (p_+, p_-). Colour ranges are scaled independently for each plot to highlight the direction of the performance value gradient for a fixed (dataset-specific) class imbalance. Contour lines are drawn at 10, 30, 50, 70, and 90% of the colour range.

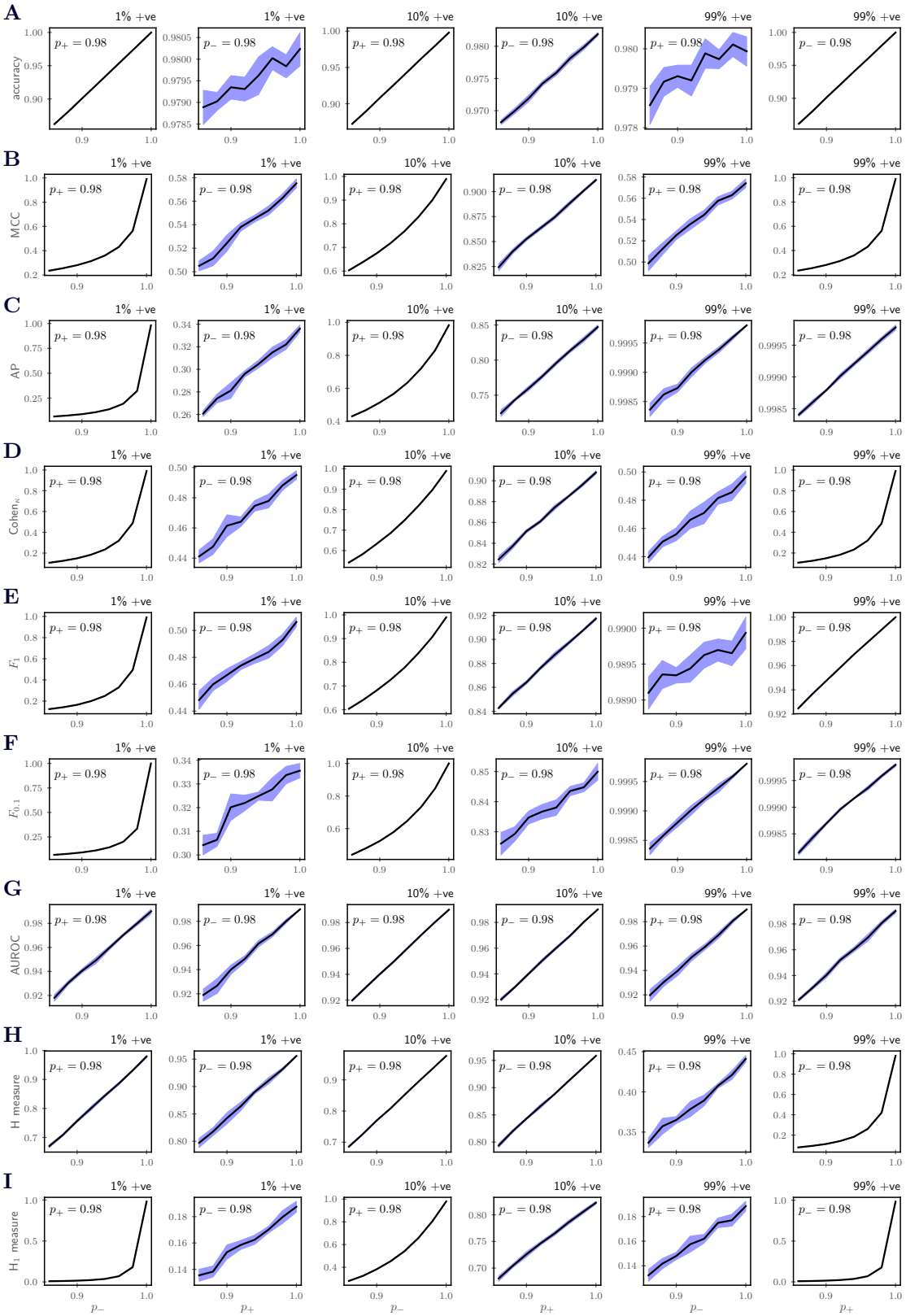


Figure 5.4.: Uncertainty in performance values due to probabilistic test conditions. Each row shows the expected average performance value (lines) and the variability due to stochastic fluctuations in the predictions (blue hull) for a given metric. The columns are sorted by class imbalance, each column showing the performance values as a function of p_+ with fixed p_- or the reverse. 95% confidence intervals are calculated from 9 simulations. For details see section 5.3.2.2.

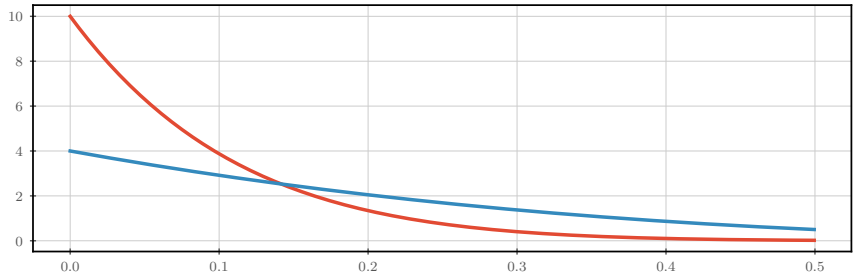


Figure 5.5.: Probability density functions of the two rank-preserving noise distributions.

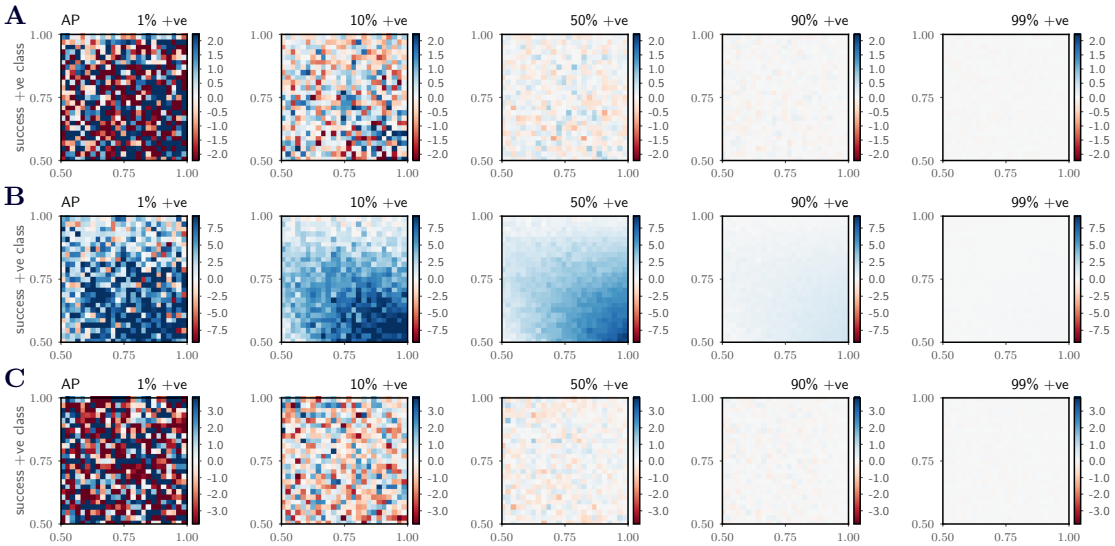


Figure 5.6.: Percentage deviation of the AP score between two simulations (relative to the first simulation). A: both simulations were conducted without added rank-preserving noise, B: the first simulation without, the second with rank preserving label noise, C: both simulations with noise, drawn from either of the rank-preserving label noise distributions shown in fig. 5.5. White represents no difference between both simulations. AP values are less reliable on probabilistic labels than on binary labels (compare colour ranges in fig. 5.6A and fig. 5.6C). Coherently blue areas in fig. 5.6B indicate a systematic change in performance value. AP is biased when it is used to compare binary labels with probabilistic labels, especially on imbalanced data.

5.3.3. Conclusion

> "It can not be emphasized enough that no claim whatsoever is being made in this paper that all algorithms are equivalent in practice, in the real world." [Wolpert, 1995]

The experiments performed here mostly serve as a lookup table for performance comparison and were designed to give an interpretable meaning to performance values. When evaluating a classifier for a specific task that has a known cost of misclassifying samples from both classes, the choice of metric is determined by the application.

The MAQC-II initiative [Shi et al., 2010], recommends using the AUROC and the MCC. The former is independent of the class imbalance, which has the advantage that its values can be compared across datasets. On very imbalanced datasets and for high-performing classifiers, the MCC is, similarly to accuracy, relatively insensitive to the minority class but is more reliable for performance ranking with test noise.

AUROC is equally sensitive to performance on both classes, insensitive to class imbalance, and produces relatively reliable rankings. The H-measure is more reliable than the AUROC for ranking performance based on the majority class but less reliable for the performance on the minority class. The above simulations are not able to determine whether H-measure is more consistent than AUROC but theoretically it should be preferred when a prior distribution over the misclassification cost can be defined [Parker, 2013]. However, using the two suggested cost weightings [Hand, Anagnostopoulos, 2014], H-measure values are not directly comparable across datasets with different class imbalances.

5.4. Background: Nonparametric performance estimation

> "That there is nothing in any object, considered in itself, which can afford us a reason for drawing a conclusion beyond it; and, That even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience;" David Hume, in A Treatise of Human Nature, book I, part III, section XII

Assessing the generalisation performance of a classifier on a sample that intersects the training data would produce overly optimistic results. A valid approach to obtain an unbiased estimate of the performance on unseen data is to use a hold out data set [Dougherty et al., 2010]. In data splitting, all available data is separated into disjoint but similar sets [Picard, Berk, 1990]: training set, test set, and for some algorithms additionally a validation set. As the names suggest, the training set is used solely to determine the model parameters. The validation set serves for estimating the generalisation performance during training and can be used for model selection. The test set data is never used for model training or selection but only for the final evaluation of the generalisation performance to unseen data [Hastie, Friedman, Tibshirani, 2009b]. This approach has multiple downsides if the number of samples is not extremely large [Harrell, Lee, Mark,

1996; Demšar, 2006]. The training data size needs to be sufficient so that adding more data does not significantly improve the model performance, the validation and test sets need to be large enough to ensure sensitivity to the generalisation performance, and the variability introduced by the split of the data needs to be negligible.

Alternative methods to simple data-splitting are cross-validation and bootstrapping. These methods make, through repeated resampling, better use of the available data and allow estimating performance measures with low bias and variability [Steyerberg et al., 2001; Efron, 1983].

5.4.1. Cross-validation and bootstrap

Cross-validation and bootstrap are data partitioning techniques that allow non-parametric statistical investigations into model performance for unseen data. They use repeated computation instead of making assumptions about the distribution of the performance results.

Cross-validation is "the most widely used error prediction technique" [Efron, 2004] and there are many different variants that differ in the way they create data splits, the level of nesting, and number of repetition [Arlot, Celisse, 2010]. In essence, all methods involve training a model on a subset of the data ("fold") and evaluating it on the hold-out sample. This is repeated for different splits of the data. Each fold yields a slightly pessimistic performance estimate if the size of the training data contributes to the model performance. By using a small test split, the bias can be reduced at the expense of higher test variability [Dougherty et al., 2010; Braga-Neto, Dougherty, 2004; Kohavi, 1995]. This can be alleviated by repeating cross-validation multiple times [Kim, 2009]. The required number of repetitions is sample size and application specific but a rule of thumb is that "more than 200 models may need to be developed and tested" [Efron, 1983; Harrell, Lee, Mark, 1996].

"[B]ootstrap procedures are nothing more than smoothed versions of cross-validation, with some adjustments made to correct for bias" [Efron, Tibshirani, 1997]. In contrast to cross-validation, bootstrap methods [Jain, Dubes, Chen, 1987; Efron, Tibshirani, 1986] sample training data from the available data using random draws with equal likelihood and replacement. Due to this sampling strategy, on average, only 63.2% of the data is unique in any training run, hence potentially reducing the diversity in the training data. The remaining data is used as test data for that bootstrap sample. The reduced training data size results in biased (pessimistic) performance estimates but the variance across bootstrap samples is reduced compared to cross-validation due to the independent sampling between boostrap estimates. This bias can be accounted for by estimating the optimism when evaluated on the training set. See [Wehrens, Putter, Buydens, 2000] for an introduction to bootstrap techniques. One of the most popular methods is the ".632 bootstrap" method, in which the true performance is estimated using a fixed weighting of the pessimistic test and optimistic training performances; in ".632+ bootstrap" the weighting is adjusted by an estimate of how much the model overfits [Efron, Tibshirani, 1997].

The obvious question, which method to choose, has been addressed many times in

the literature. Results vary presumably due to different scoring rules, different signal to noise in the data and models with varying tendencies to overfit. One of the most cited studies concludes that 10-fold cross-validation is optimal when accuracy is the criterion [Kohavi, 1995]. For logistic regression, bootstrap methods are reported to be more efficient [Steyerberg et al., 2001; Harrell, 1998]. When matched for compute-resources, repeated k-fold CV and bootstrap work equally well [Kim, 2009]. For small samples, the .632+ bootstrap can be biased for high signal to noise ratio [Molinaro, Simon, Pfeiffer, 2005] but this is not observed for larger sample sizes, where k-fold cross-validation and bootstrap perform equally well [Molinaro, Simon, Pfeiffer, 2005; Beleites et al., 2005].

5.4.2. Deep learning

While cross-validation is considered to be an important tool to assess generalisation ability in machine learning, a good choice of performance metric and statistical rigour for model comparison have been lacking for some time in machine learning practice [Demšar, 2006].

In deep learning experiments, cross-validation, analysis of the impact of hyper-parameters and hypothesis-driven model comparisons are rarely reported, which Lucic et al. attribute to the high computational cost of training multiple models and the rapid advances in the field [Lucic et al., 2017]. The size of the training data is in many deep learning vision tasks the major driving factor of algorithmic performance [Halevy, Norvig, Pereira, 2009; Sun et al., 2017; Valle et al., 2017]. It is therefore common practice to use all available data for training to leverage this additional information and to sacrifice the ability to accurately evaluate the model's generalisation performance; or worse, to optimise parameters using the test set [Valle et al., 2017].

For an account on how deep neural networks are trained in practice, see [Valle et al., 2017]. Valle et al. developed a model for melanoma detection using multi-way ANOVA to determine the effect of model hyper-parameters. Note that although one aim of that study is to "investigate methodological issues for designing and evaluating deep learning models for melanoma detection", multi-way ANOVA is questionable in its statistical validity for classification comparisons (and underpowered) [Demšar, 2006], their experimental design does not account for the sample selection variability using bootstrap or cross-validation techniques, and p-values appear not to be corrected for multiple comparisons. Also, a comparison of models that are state of the art with models found using ANOVA is confounded by the larger training data in the latter case with "more careful ... matching of diagnostics among the sources" likely increasing the data quality as well.

Splitting the training data into k-fold, cross-validation might degrade classifier performance substantially leading to different design decisions [Lu et al., 2016]. Furthermore, algorithm design decisions such as degree of data augmentation or regularisation depend on the size of the training data and can be difficult to estimate using cross-validation especially for small datasets. Cross-validation can introduce problems if used inappropriately. Multiple classifiers learned via cross-validation are not independent [Dietterich, 1998] and cross-validation folds do not share the same data distributions. This can introduce biases in statistical estimates [Forman, Scholz, 2010]. For binary classification using

F-score and AUROC, the type of cross-validation and its exact implementation can cause biases which become especially relevant in the case of class imbalance [Forman, Scholz, 2010]. For the calculation of the area under the precision recall curve (average precision), Boyd, Eng, Page conclude that bootstrapping and 10-fold cross-validation produce unreliable estimates in skewed or imbalanced datasets and that they require large number of estimates and therefore computation [Boyd, Eng, Page, 2013].

Also, cross-validation schemes can not account for human biases such as proficiency. Shi et al. performed a large-scale analysis across 36 teams and more than 30'000 models trained on the same dataset but applied to different prediction tasks to test generalisation performance. The prediction task was the most important factor in the algorithm design but also proficiency of the researchers played an important role for the generalisation performance.

5.4.3. Conclusions

Cross-validation and bootstrap methods require prohibitively large numbers of training iterations for deep learning to estimate model performance with low bias and variance. However, if model selection or ranking is the goal and if bias can be assumed to be consistent across models, then sampling methods with low variability but higher bias (such as bootstrap) are preferable [Kohavi, 1995]. For instance, Paass uses 30 bootstrap samples for model selection in neural network design [Paass, 1993]. However, the performance bias using bootstrap methods is data-dependent, making it harder to compare methods across datasets. Despite warnings against benchmarking methods [Drummond, Japkowicz, 2010], reporting pessimistic performance results might not be in the interest to researchers trying to beat the state of the art.

Alternatively, methods can be compared across domains to test their generalisability [Demšar, 2006], but this is difficult for algorithms that are specialised to a domain, especially if algorithms require large amount of data, as the case with deep learning in medical imaging. To conclude, in deep learning, performance and validation are both limited by the size of the data and compute resources and in practice emphasis is placed on improving performance. This is at the expense of higher uncertainty about the validity of the results, possibly creating an "illusion of progress" [Hand, 2006].

Part II.

Results

Chapter 6

Motion artefact classification using convolutional neural networks

Contents

6.1. Introduction

6.1.1. Motion artefact detection and correction

In diffusion weighted MRI (dMRI), diffusion weighting is achieved via large gradients that amplify the dephasing and therefore signal attenuation due to random movement of protons. As reviewed in section 3.2, this makes dMRI sensitive to bulk motion. The most common approach to reduce bulk motion sensitivity [Le Bihan et al., 2006] is an imaging technique called single-shot Echo Planar Imaging (EPI), which acquires all k-space data required to reconstruct a slice rapidly after a single excitation and applied diffusion weighting. A volume is typically acquired as a series of parallel 2D slices, acquired with interleaved ordering (see section 3.2.4.5).

Depending on the duration of a subject's movement, motion artefacts affect isolated or multiple slices within a volume and can extend across multiple volumes. In adults, motion is often slow and can mostly be accounted for by correcting the location and orientation of the acquired volumes through image registration techniques [Leemans, Jones, 2009; Zwiers, 2010].

Incoherent subject movement during the diffusion preparation can cause phase errors that induce signal loss on the slice-level. Motion during the acquisition of the volume introduces misalignment of acquired slices, most prominently visible as a mismatch between adjacent slice positions, and signal dropout due to spin history effects.

Recent work has addressed the within-volume misalignment through slice to volume registration and aims at detecting artefacted data through comparison with data acquired in adjacent locations in q-space [Andersson, Sotiropoulos, 2015b].

Motion corrupted volumes can decrease the quality of registration-based pre-processing steps such as the correction of susceptibility and eddy-current distortions, and lead to biased scalar diffusion measures and fibre directions [Pannek et al., 2012a]. Quality control and artefact detection, correction and robust fitting procures are typically incorporated at multiple stages in the diffusion data processing pipeline and there are many techniques that aim to robustly represent the diffusion signal and/or derive artefact-free diffusion model parameters [Mangin et al., 2002; Chang, Jones, Pierpaoli, 2005; Andersson, Sotiropoulos, 2015a]. See [Liu, Zhu, Zhong, 2015] for a comparison of 3 software packages geared towards artefact detection and correction in diffusion tensor imaging. Here, I focus on the very first stages of the processing pipeline, where motion artefact detection is applied to the reconstructed raw diffusion images, not the diffusion model parameters.

There are a variety of motion artefact detection tools that use residuals of diffusion tensor fits [Chang, Jones, Pierpaoli, 2005], residuals of higher-order representation fits [Pannek et al., 2012b] or smoothness constraints in the q-space domain [Andersson et al., 2016] to reject or correct corrupted data on the slice and voxel level. Motion artefact correction methods rely on the redundancy of uncorrupted data through repeated scans or adjacency in q-space.

In adults, severe intra-volume motion affects only about 1.4% of the data [Ling et al., 2012] but in neonatal imaging, motion is much more prevalent and severe, requiring additional focus on quality control [Pannek et al., 2012a; Heemskerk et al., 2013]. From local experience, the artefact detection and replacement of the software tool eddy [Andersson et al., 2016] (without slice to volume registration) can cope with a small number of severely motion affected volumes but produces artefacts in the presence of many motion corrupted volumes, an observation confirmed at other sites working with neonates. Studies on neonates report the exclusion of 13 to 26% of volumes prior to processing [Pavaine et al., 2016; Dudink et al., 2007; Pul et al., 2012; Van Kooij et al., 2011].

Finding subtle artefacts in diffusion images requires experience but it is easy for a human to spot severe motion corruption. Common practice is to remove severely motion corrupted volumes manually and use software to process the remaining data. However, manual removal of severely motion-corrupted volumes is time-consuming and might be subjective. There are a number of approaches to detect and correct for motion artefacts in dMRI data. At image resolutions of $1mm$ or coarser, sharp and large steps in intensity between adjacent slices are likely not of anatomical origin but caused by motion artefacts. A simple to implement quality control measure is to quantify inter-slice intensity variations [Li et al., 2013]. However, this technique requires a manually set threshold that is likely dependent on the size of the brain and the b-value and the authors did not test their method on b-values above $1000mm/s^2$. Pannek et al. take a registration-based approach by dividing a diffusion volume into its odd and even slices and registering

those rigidly to estimate the within-volume motion. The Frobenius norm[1] of the rigid transformation matrix and the identity matrix serves as a motion severity measure. This method has the drawback that it ignores signal dropout and the Frobenius norm is likely a poor proxy for assessing artefact severity as changes of the rotation matrix entries are weighted equally to those of the translation vector measured in real-world coordinates.

Motion artefacts can also be detected by extracting information from the local image texture. In [Liu et al., 2013], motion artefacts are detected and corrected by fitting smooth within-slice image intensity profiles and measuring intra-volume motion and contrast artefacts by analysing image intensity correlations. Zhou et al. use local binary patterns, which is a texture analysis method used in computer vision (see [Bouwmans et al., 2016] for a review), to account for motion artefacts in diffusion tensor fitting [Zhou et al., 2011].

All approaches mentioned require some form of calibration to form a decision about usability or to weight the contribution of the data. Depending on the extracted features, this calibration can depend on scanning or anatomical parameters such as b-value and the size of the brain, requiring re-calibration for the specific target application, or on the subject-level. An alternative to designing feature extraction methods that indicate motion is to train a supervised classification algorithm to do both; extract features and calibrate them at the same time. This requires ground truth labels generated via simulations of motion or human-generated annotations.

Recently, [Lorch et al., 2017] employed a supervised machine learning approach using decision forests to detect texture specific to motion in T_2-weighted images with motion simulated in k-space. We applied deep convolutional neural networks, which are the state of the art technique in image classification [Krizhevsky, Sutskever, Hinton, 2012], to classify diffusion-weighed volumes into usable and severely motion-corrupted, matching human-generated annotations [Kelly et al., 2017].

6.1.2. Motion artefact detection using neural networks

This chapter addresses questions on how to design and train an automated, fast and data-driven neural network classification algorithm for the assessment of severe motion artefacts in neonatal diffusion images with the aim to find and train a neural network architecture that can replace the manual task of removing severely motion artefacted ('outlier') volumes. This task is usually performed by a human, which makes it a subjective, labour intensive and possibly inconsistent process. I use artificial neural networks to learn to distinguish severely motion corrupted volumes from volumes that I labelled as either usable or outlier. For background about artificial neural networks in the context of image classification see section 4.3.

An alternative approach to generate ground truth labels is to simulate motion artefacts [Drobnjak et al., 2006; Lorch et al., 2017], which has the advantage of having control over the severity and quality of motion and is less labour intense than manual data annotation. However, the quality of the classifier would depend on how realistic and

[1]The Frobenius norm is the square root of the sum of the squared elements of a matrix.

applicable the simulator is for the target application. By using human-generated labels to train a neural network that can generalise these decision criteria to unseen data, it is possible to remedy the labour intensity and inter-operator variability. However, it assumes that the annotations are of sufficient quality to improve downstream diffusion image processing and analysis.

It would be ideal to exclude the human decision making in the first place, and instead build a pipeline that removes outlier volumes to directly improve the outcome of the diffusion analysis. This could be achieved, for instance, by reinforcement learning or by optimising a neural network that performs the full processing pipeline. Besides a high computational cost, this is only possible if a suitable and robust metric for assessing the output of the diffusion analysis can be defined. In this work, I use human judgement of image quality as a surrogate for directly optimising the performance of the diffusion analysis pipeline and focus on answering the questions:

- Can we build a neural network that classifies outlier volumes with near human-level performance?
- How much training data is required and what are good ways to sample and augment data?
- How does the within-class structure of the diffusion data affect performance?
- Is it possible to adapt existing neural networks trained on unrelated computer vision classification tasks to perform motion artefact detection with less training data or with superior performance?

Related work The successful application of deep neural networks to general vision problems in recent years has prompted researchers to investigate their applicability to answering medical questions. The main challenge is that medical datasets are more difficult to collect and are therefore typically much smaller and often imbalanced in their class composition compared to curated general vision datasets. However, networks trained on these large datasets have a surprising versatility and performance on unrelated datasets and vision tasks that they have not been trained for. The technique of reusing parts of these networks for unsupervised feature extraction or retraining the last layers to a new dataset is referred to as transfer learning and is commonly used in medical image classification [Yosinski et al., 2014; Sharif Razavian et al., 2014].

In the context of lymph-node detection and lung-disease classification in CT images, Lu et al. find that deep vision models pre-trained on large general images provide a good initialisation which, in the case of sufficient target domain training data, can be substantially improved with fine-tuning using medical datasets [Lu et al., 2016]. In many applications, deep models perform better than shallow models if enough training data is available and pre-trained models fine-tuned for the target domain, often outperform models trained from scratch [Lu et al., 2016].

In general, deep learning profits from better models and more and better training data. While an increased data size has allowed remarkable progress in classification accuracy, improvement on fixed datasets is important ongoing research [Zhu et al., 2012], especially for medical imaging applications, where the data size is commonly extremely

limited [Cho et al., 2015]. In [Valle et al., 2017], the authors perform an exhaustive search to find parameters that yield the best performing melanoma screening models. They investigate different model architectures, pre-processing methods and the effect of data-augmentation for training and test sets.

However, melanoma classification and lung disease classification are rather complex computer vision problems that are similar to general image classification tasks and require comparatively large contextual information using large receptive fields and deep networks and it is not clear if these findings translate to the domain of dMRI motion artefact detection.

6.1.3. Neural network architecture

The neural network architecture used here is inspired by and adapted from the popular VGG network architecture [Simonyan, Zisserman, 2014]. It is named after the Visual Geometry Group (University of Oxford) that submitted the model for the ImageNet competition ILSVRC-2014. It was designed to classify colour photos of humans, animals, plants, buildings, and general objects of the imagenet dataset [Deng et al., 2009] into 1000 categories.

A number of architectures exist that perform similarly or even better on the imagenet dataset and some have less parameters to train. The reasons I opted for this model are its relatively high accuracy in object detection tasks, the availability[2] of network weights pre-trained on the ImageNet challenge dataset, and its relatively simple structure, which simplifies hyper-parameter optimisation and facilitates a future implementation as a standalone image processing tool. However, the optimal choice or the automated design of network architectures for a specific task is an open research question. See section 4.3 for a brief discussion of building blocks of convolutional neural networks.

The VGG network uses convolution layers with the smallest possible 2D filter size (3x3) that are applied with a step size (stride) of 1. By stacking multiple convolution layers without pooling their output, it is possible to increase the effective receptive field of each unit with a smaller number of parameters compared to a network that uses a single convolution layer instead and filters that cover the same receptive field [Simonyan, Zisserman, 2014]. Furthermore, convolution layers use rectified linear activation units, which allows learning non-linear functions, whereas single layers with large convolution filters do not have this degree of freedom. These blocks of convolution layers are separated by 2x2 maximum pooling layers with stride 2 and followed by a small number of fully connected layers at the end of the network. Convolution layers in each block have equal spatial dimensions but the stride 2 local pooling layers reduce the spatial dimensions of the network in a step-wise fashion with increasing depth (see table 6.25).

The input of the VGG16 network is 3 dimensional with the 3 colour channels as the third dimension and it has 64 convolution filters in the first layer. The name of the network indicates that it consists of 16 fully connected layers. These layers are arranged in 5 blocks consisting of 2, 2, 3, 3, and 3 convolution layers, followed by a flattening

[2] http://www.robots.ox.ac.uk/~vgg/research/very_deep/

operation that converts the 2D feature map into a 1D vector through concatenation. The last 3 layers are fully connected ('dense') layers, which compute a nonlinear combination of the extracted features, and finally yield a vector representing the class correspondence likelihood.

However, the number of convolution filters doubles for each of the first 4 blocks, which leads to an increase in the parameters of the convolution layers from 1792 in the first layer to 2.36 million parameters in the last convolution layer. The first dense layer has 102 million parameters, which is by far the highest parameter count of any layer.

Table 6.1.: The original VGG16 network architecture.

	name	activation	output shape	parameters
1	InputLayer		(224, 224, 3)	0
2	Conv2D 3x3	relu	(224, 224, 64)	1792
3	Conv2D 3x3	relu	(224, 224, 64)	36928
4	MaxPooling2D 2x2		(112, 112, 64)	0
5	Conv2D 3x3	relu	(112, 112, 128)	73856
6	Conv2D 3x3	relu	(112, 112, 128)	147584
7	MaxPooling2D 2x2		(56, 56, 128)	0
8	Conv2D 3x3	relu	(56, 56, 256)	295168
9	Conv2D 3x3	relu	(56, 56, 256)	590080
10	Conv2D 3x3	relu	(56, 56, 256)	590080
11	MaxPooling2D 2x2		(28, 28, 256)	0
12	Conv2D 3x3	relu	(28, 28, 512)	1180160
13	Conv2D 3x3	relu	(28, 28, 512)	2359808
14	Conv2D 3x3	relu	(28, 28, 512)	2359808
15	MaxPooling2D 2x2		(14, 14, 512)	0
16	Conv2D 3x3	relu	(14, 14, 512)	2359808
17	Conv2D 3x3	relu	(14, 14, 512)	2359808
18	Conv2D 3x3	relu	(14, 14, 512)	2359808
19	MaxPooling2D 2x2		(7, 7, 512)	0
20	Flatten		(25088,)	0
21	Dense	relu	(4096,)	102764544
22	Dense	relu	(4096,)	16781312
23	Dense	softmax	(1000,)	4097000

I used the Keras deep learning library (version 2.0.8) with the Tensorflow 1.2.1 backend and scikit-learn 0.19.0, pandas 0.20.2 and R 3.4.1 with the package hmeasure 1.0 for the statistical analysis of all diffusion image artefact classification experiments. For the fashion image classification experiments, I used Tensorflow version 1.4.0 and its incorporated version of Keras.

6.2. Effect of class imbalance on image classification

This section explores the effect of class imbalance on the performance of a typical computer vision classification problem using the VGG model architecture (see section 6.1.3). The aim of the following simulations is to investigate the influence of reduced training size, within-class and between-class imbalance on the classification performance and whether sample weighting and oversampling aid in training classifiers on skewed data that perform well on an unskewed test dataset.

One way to simulate within-class structure for binary classification is to group a dataset

that contains multiple categories into two meta-categories ("groups"). By comparing neural network performances of models that were trained to distinguish the original categories with networks that had only access to the meta-categories, it is possible to assess the value of learning the within-class structure.

For the application of motion artefact detection, within-class imbalance can consists of types of artefacts that manifest differently at different b-values. Due to the higher number of b=2600 images, those artefacts would be overrepresented features if all data was fed to the network in the same proportion as present in the training set.

6.2.1. Data

I use a publicly available dataset of fashion images [Xiao, Rasul, Vollgraf, 2017] that contains 60 000 training images (5 000 of which were exclusively used for evaluation during training) and 10 000 test images. I chose this dataset for practical reasons: it contains 10 classes with varying degree of between-class similarity, is composed of relatively small grayscale images (28x28 pixels), which allow fast training, and it is a more challenging dataset than for instance handwritten digit recognition, which is considered by some researchers as too easy to test modern neural networks on [Perez, Wang, 2017]. An example selection of images from each category is shown in figure 6.1.

Figure 6.1.: Exemplary samples of the fashion dataset of the categories 0 to 9 from left to right.

To assess binary classification performance, two classification target groups are defined as categories 0 to 4 and categories 5 to 9 (see fig. 6.1). By varying the number of training

category	*full*	*small*	*02*	*02_13*
0 to 4	0.500	0.300	0.500	0.500
5 to 9	0.500	0.301	0.101	0.101
5	0.100	0.060	0.020	0.010
6	0.100	0.061	0.020	0.015
7	0.100	0.060	0.020	0.020
8	0.100	0.060	0.020	0.025
9	0.100	0.060	0.020	0.030

Table 6.2.: Sample size fractions of different classes (0 to 9) relative to the full training data set (n=55000 images) for each partition. Binary classes are categories 0 to 4 and categories 5 to 9. Note that the within-class imbalance in the *02_13* fraction affects only the minority class.

samples in each category, it is possible to assess the effect of data size, within-class imbalance, and of between-class imbalance. For this, I use 4 different partitions: the partition *full* contains the complete training dataset and the other 3 partitions contain 40% less data, without group imbalance (*small*), with 5:1 between-group imbalance (*02*), and with additional within-group imbalance in the smaller group (*02_13*). The fractions of training samples in each partition are listed in table 6.2 and plotted in figure 6.2.

All models were validated during training and finally tested on the same distinct (and unskewed) validation and test datasets. All models were trained 4 times on randomly generated training partitions with random selection from the training data pool (bootstrap, [Efron, Tibshirani, 1986]) to obtain an estimate of the variability of model performance.

For each individual bootstrap sample of a partition, images were sampled without replacement but samples were drawn with replacement between bootstraps. This ensures that the model fitting on categories with fewer examples are not dominated by particular examples but are representative of the category. Note that samples can be part of multiple bootstraps and that the number of unique samples depends on the relative size of the respective categories. The full partition bootstraps differ only in the order of samples presented to the models.

In contrast to typical non-parametric bootstrapping, for each partition, data were sampled without replacement. This is motivated by two factors. First, random sampling with replacement would decrease the variability in the fully sampled partition effectively decreasing the sample size making the results harder to interpret. Second, theoretical findings on different but possibly related problems (strongly convex loss functions [Gürbüzbalaban, Ozdaglar, Parrilo, 2015], least means squares optimization problems [Recht, Re, 2012]) show that higher learning rates are expected for randomly reshuffled data compared to data drawn with replacement [Gürbüzbalaban, Ozdaglar, Parrilo, 2015].

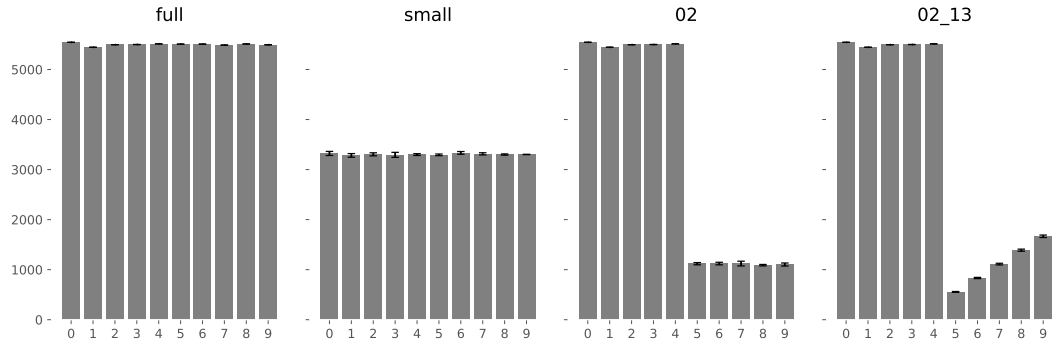


Figure 6.2.: Distribution of sample size split by category for different training data partitions. Error bars indicate one standard deviation due to bootstrap sampling.

6.2.2. Network architecture

The vision model architecture is an adapted and extended version of [Rasul, 2017]. It is a typical VGG-inspired vision network and contains 483'682 trainable parameters. It consists of two convolution blocks followed by a block of three fully connected layers (see table 6.3). All convolution and fully connected layers use rectified linear activation functions. The convolution blocks contain two 2D convolution layers, each with 32 filters in the first and 64 filters in the second block. The convolution blocks are separated by maximum pooling layers, with pooling size 2x2 and stride 2, followed by a dropout layer (see section 4.4.1) with dropout rate 0.25 to prevent overfitting. The output of the second max-pooling layer is fed into a batch normalisation (see section 4.4.1) and another dropout layer with dropout probability 0.25. The majority of the model's parameters are in the fully connected layer mapping the 3136 features to 128 features.

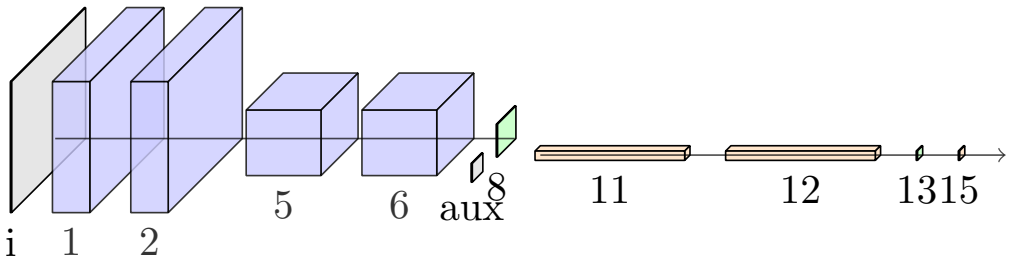


Figure 6.3.: Convolution and dense layers of the hybrid model consisting of the vision model and auxiliary information layer that are concatenated and used as input for layer 11. Input layers are shown in grey (input image i and auxiliary input aux), convolution layers are shown in purple, batch normalisation in cyan and dense layers in orange. The height and depth of the boxes represents spatial extent, the width represents number of features. Omitted are dropout, pooling, flatten and concatenate layers. Layer numbers refer to tables 6.3 and 6.5.

	name	activation	input shape	output shape	parameters
1	Conv2D 3x3	relu	(28, 28, 1)	(28, 28, 32)	320
2	Conv2D 3x3	relu	(28, 28, 32)	(28, 28, 32)	9248
3	MaxPooling2D 2x2		(28, 28, 32)	(14, 14, 32)	0
4	Dropout (p=0.25)		(14, 14, 32)	(14, 14, 32)	0
5	Conv2D 3x3	relu	(14, 14, 32)	(14, 14, 64)	18496
6	Conv2D 3x3	relu	(14, 14, 64)	(14, 14, 64)	36928
7	MaxPooling2D 2x2		(14, 14, 64)	(7, 7, 64)	0
8	BatchNormalization		(7, 7, 64)	(7, 7, 64)	256 (128)
9	Dropout (p=0.25)		(7, 7, 64)	(7, 7, 64)	0
10	Flatten		(7, 7, 64)	(3136,)	0
11	Dense	relu	(3136,)	(128,)	401536
12	Dense	relu	(128,)	(128,)	16512
13	BatchNormalization		(128,)	(128,)	512 (256)
14	Dropout (p=0.5)		(128,)	(128,)	0
15	Dense	softmax	(128,)	(2,)	258

Table 6.3.: Vision model architecture for classification into two categories using 484066 parameters 384 of which are non-trainable. For each layer, the total number of parameters is shown. The batch size dimension is omitted. Note that shape tuples reflect the dimensionality of the internal representation, i.e. (128,) is a 1D vector, (28,28,1) is a 3D array consisting of one 2D slice. If a layer has fixed or non-trainable parameters, they are shown in brackets next to the total number of parameters. "relu" stands for rectified linear unit (see section 4.3). The shape of most layers is illustrated graphically in fig. 6.3.

notation	description
10	trained on 10 classes
10+10	trained on 10 classes, fully connected layers retrained on 10 classes
10+bin	trained on 10 classes, fully connected layers retrained on 2 classes
bin+10	trained on 2 classes, fully connected layers retrained on 10 classes
bin	trained on 2 classes
bin+bin	trained on 2 classes, fully connected layers retrained on 2 classes
(aux&vis)bin	trained on 2 classes with auxiliary information input
(aux&vis)bin+bin	as bin(aux) but fully connected layers retrained without auxiliary information
o	between-group balanced over-sampling
os	between-group and within-group balanced over-sampling
w	weighting to balance the between-group imbalance
ws	weighting to balance the between and within-group imbalance

Table 6.4.: Notation of models and sampling schemes.

	name	activation	input shape	output shape	parameters
aux	Dense (auxiliary)		(1,)	(1,)	2
vis	Vision (layers 1-10)		(28, 28, 1)	(3136,)	65248 (128)
10b	Concatenate (aux & vis)		[(1,), (3136,)]	(3137,)	0
11	Dense	relu	(3137,)	(128,)	401664
12	Dense	relu	(128,)	(128,)	16512
13	BatchNormalization		(128,)	(128,)	512 (256)
14	Dropout		(128,)	(128,)	0
15	Dense	softmax	(128,)	(2,)	258

Table 6.5.: Combined vision and auxiliary information model for binary classification. The vision model architecture is identical to the first 10 layers in table 6.3. The auxiliary input is concatenated to the output of the vision model. The combined model has 484196 parameters, 384 of which are non-trainable (shown in brackets). For a graphical representation see fig. 6.3.

Models labelled *bin* are trained to perform a binary classification task of deciding whether images belong to the first five or last five categories (columns in figure 6.1). The first group contains only clothing articles whereas the second group contains shirts and 4 accessory categories. Models labelled *10* were trained to recognize all categories, which is a more challenging classification problem but also potentially allows the model to learn better feature representations [Zhu et al., 2012; Esteva et al., 2017].

To investigate the performance of the vision network when presented with all 10 training labels on the binary classification task, I convert the output of the *10* models to the binary classification task by changing the prediction from the most likely category to the corresponding group.

Most experiments are performed using the vision architecture shown in table 6.3. A hybrid model that uses the images and an additional input of size 1 is shown in table 6.5. The output of the features after 10 layers of the vision network is combined with this additional information layer and fed combined to the last 3 classification layers. This allows passing "auxiliary" information to the network that can be combined with the vision-based features. These networks are denoted as *(aux&vis)bin*. Non-image auxiliary information can improve classification performance if it is relevant for the classification task. In the context of motion classification, auxiliary information could be the b-value, information about the slice position, or the subject's age or gender. Furthermore, this information can be used to indirectly guide the network's learning process even if auxiliary information is not present during inference as discussed below.

6.2.3. Training

Each model was trained using the Adam optimiser [Kingma, Ba, 2014] for 12 epochs without data augmentation (see section 4.4.1) followed by at least 100 epochs with data augmentation (0 to 8 degree rotation, 0 to 10% translation, both uniformly sampled) or until the model's validation loss (cross-entropy) did not decrease for more than 12 epochs. The latter was the case for one data-split, bootstrap, model and sampling method combination (*02*, *10*, *w*) and took 5 more epochs to converge. The Adam parameters are

$\beta_1 = 0.9$ and $\beta_2 = 0.999$, no learning rate decay and a learning rate of 0.001. The gradient moment estimates were reset between training stages and all models were trained with a batch size of 128 images.

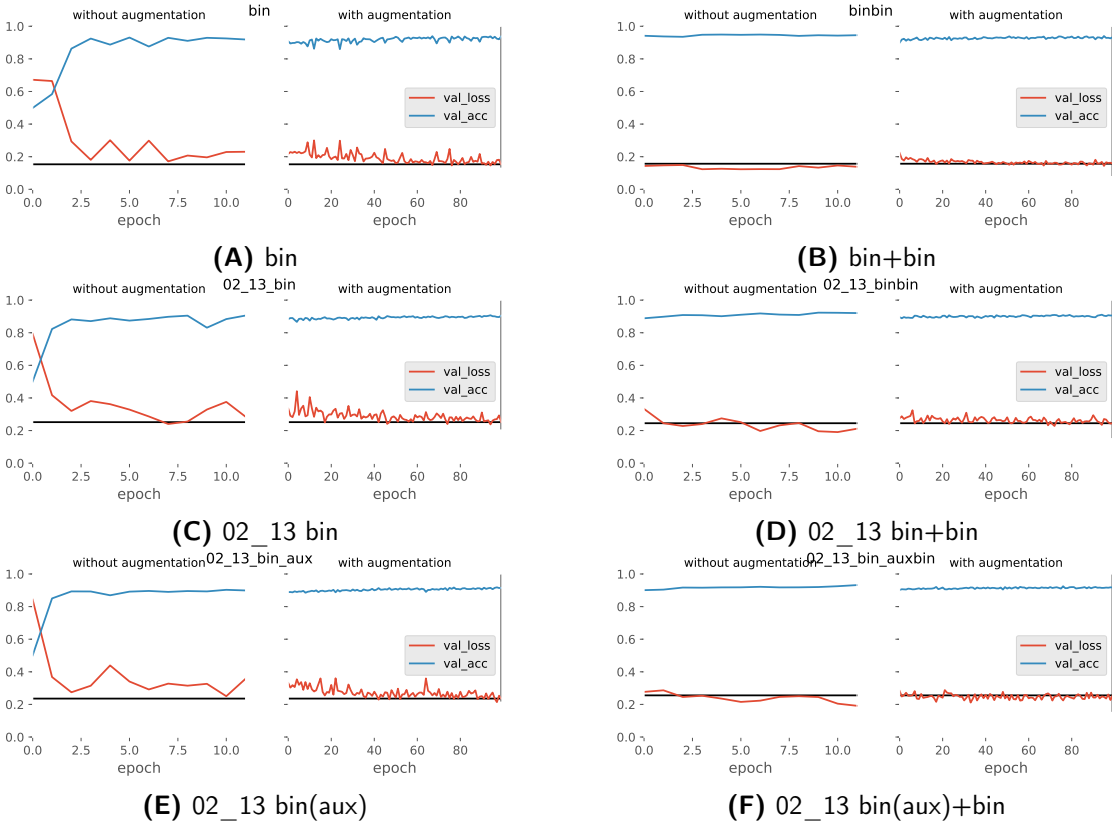


Figure 6.4.: Exemplary learning curves for the binary classification problem on the full dataset (top) and the 02_13 split (middle and bottom). Blue curves show the validation accuracy, red curves the validation loss (cross-entropy). The left columns show the training progress of the full network, the right columns display the progress of the networks on the left retrained after resetting the final layers (number 11 to 15).

During training, each model's state was saved if its validation loss was an all-time low. For testing, the model with the lowest validation loss was used unless a model from a later epoch had a validation loss that differed from the best validation loss by less than 0.05. Exemplary validation loss and accuracy curves are shown in fig. 6.4. This ensures that models were exposed to similar number of epochs irrespective of their convergence rates while providing a failsafe for models that overfit. The choice of data augmentation scheme is inspired by public Keras code that classifies similar data using VGG-style architectures.

Following this two-stage training, the last classification layers after the flatten layer were reset to random weights and trained on the same data again, repeating the 2-stage

process described above. The lower layers (1 to 9) were kept constant during this second training round. Models that were partially re-trained have a *+* sign in their name. For instance, the model *10+bin* was first trained on 10 categories, stripped of its last layers, and received randomly initialised classification layers, which were trained to perform binary classification.

Class imbalance can have detrimental effects on classifier performance (see section 4.4.2). Two methods to improve training on imbalanced data are data sampling and cost weighting according to the distribution of samples. I use minority class oversampling, denoted as *o*, to account for the imbalance between-groups. Each training batch contains an equal number of samples from each category, hence in a training epoch some or all samples from the minority class are used multiple times (but with different random transformations).

Another approach to account for class imbalance is to weigh the misclassification cost of minority samples higher than that of majority samples. I use the weighting scheme proposed in [King et al., 2001] for logistic regression, which weights the samples with binary or multi-class ground truth labels $\mathbf{l}$ based on their prevalence:

$$w(l) = \frac{N_{\text{samples}}}{2\sum_{i=1}^{N_{\text{samples}}} I(\mathbf{l}_i = l)} \tag{6.1}$$

$$w_s(l) = \frac{N_{\text{samples}}}{10\sum_{i=1}^{N_{\text{samples}}} I(\mathbf{l}_i = l)} \tag{6.2}$$

with $I(x)$, the indicator function.

In the *02_13* partition, the minority class also has a skewed within-class distribution. Assuming that one knows the within-class structure, it is possible to also account for this in the weighting and sampling methods of the binary classifier. The variants of sampling and oversampling that use the multi-class labels are denoted as *ws* and *os*, respectively.

The hybrid models *(aux&vis)bin* are trained to perform binary classification, similarly to the *bin* models. However, during training (and testing), this hybrid model has access to the additional auxiliary information, whether the sample is from an odd or even class (parity). This information is fed into the network after the first 10 layers of the vision architecture and therefore contributes to the training of the last dense layers, which in turn might have an indirect effect on the lower vision layers as both are trained jointly via back-propagation.

To test how auxiliary information affects the training of the first vision layers, this model is stripped of its auxiliary input and the last classification layers, and receives randomly initialised binary classification layers that are retrained for binary classification in the same way as the *bin+bin* networks. This model, denoted *(aux&vis)bin*, has the same architecture as *bin* (or *bin+bin*) and was trained for the same number of epochs as the other two-stage models.

See table 6.4 for a complete summary of the model and data sampling notations used. For models that were trained twice, the same sampling (or weighting) scheme were used throughout. Not all combinations of models and sampling strategies were explored due to time-constraints.

6.2.4. Results and discussion

All performance results are reported as average and 95% confidence interval across 4 models trained independently on different bootstrap samples in the form mean [lower bound, upper bound]. The 95% confidence intervals are calculated using the t-distribution estimate on the logit-transformed measures. Due to computational constraints[3], models were trained and evaluated on only 4 bootstrap repetitions, which presumably causes confidence intervals that cover more than 95% of the true variability.

All results report AP and AUROC scores to represent one metric typically used in information retrieval and in classifier evaluation. Note that AP and AUROC scores in the multi-class models refer to the performance on the binary task, which is assessed by transforming the prediction label, not to the multi-class AP or AUROC, usually calculated by averaging the (weighted) scores across labels.

6.2.4.1. Balanced dataset

The *full* partition contains the complete training set and all classes are equally represented. This serves as the baseline partition as it represents ideal conditions to train the classification networks. The vision network achieves an AUROC of 0.9905 on the binary problem, when trained to classify all 10 clothing categories (*10* in table 6.6). The model trained on the binary classification directly, achieves a lower AUROC of 0.9868 (*bin*). The higher performance of the model trained on the 10-class problem shows the value of features that differentiate the within-class structure in this grouping. This effect is likely dependent on the imbalance of clothing and accessory articles in the binary grouping of the data and therefore dataset dependent.

In both cases, performance can be improved by retraining the last layers from scratch. Training the first layers on the binary problem and then retraining the last layers to classify 10 categories (*bin+10*) results in a reduced model performance compared to the *10+10* model. Hence, the first 10 layers of the *bin* networks have not learned a sufficient representation of features that allows delineating the 10 classes, which in turn degrades performance on the binary classification task.

All models that train on the 10-class problem outperform models that train on the binary problem. This is the case for training the full model, for the first layers and for the retrained last layers. Hence, learning features that delineate the within class-structure improves performance.

The *small* partition contains 40% less data than the *full* partition. The AUROC of all models drops compared to that trained on the *full* dataset: *10+10* drops from 0.9919 to 0.9907 and *bin+bin* from 0.9873 to 0.9860. However, the ranking between classifiers remains the same (table 6.7).

[3]Training of all models took approximately one week on a Nvidia GeForce GT 730M GPU.

classifier	sampling	AP	AUROC
10		0.9899 [0.9878, 0.9916]	0.9905 [0.9888, 0.9921]
10+10		0.9914 [0.9897, 0.9928]	0.9919 [0.9903, 0.9932]
10+bin		0.9910 [0.9895, 0.9922]	0.9915 [0.9902, 0.9926]
bin+10		0.9859 [0.9853, 0.9865]	0.9871 [0.9868, 0.9874]
bin		0.9855 [0.9840, 0.9869]	0.9868 [0.9855, 0.9879]
bin+bin		**0.9861** [0.9844, 0.9876]	**0.9873** [0.9860, 0.9885]

Table 6.6.: Results for the *full* partition split by models that had access to all 10 categories are shown on the top. The best average performance for the binary models is highlighted in bold font. AP stands for average precision, AUROC for area under the receiver operator curve. See table 6.4 for classifier and sampling notation.

classifier	sampling	AP	AUROC
10		0.9896 [0.9880, 0.9910]	0.9903 [0.9889, 0.9916]
10+10		0.9901 [0.9886, 0.9914]	0.9907 [0.9894, 0.9919]
10+bin		0.9895 [0.9875, 0.9913]	0.9904 [0.9887, 0.9918]
bin+10		0.9846 [0.9785, 0.9890]	0.9859 [0.9805, 0.9898]
bin		0.9831 [0.9748, 0.9887]	0.9849 [0.9782, 0.9896]
bin+bin		**0.9846** [0.9765, 0.9899]	**0.9860** [0.9792, 0.9906]

Table 6.7.: Results for the *small* partition split. See table 6.4 for classifier and sampling notation.

6.2.4.2. Between-group imbalance

In the *02* partition, the categories 5 to 9 are reduced by 80% compared to the *full* partition but each category is equally represented. The ratio of the first group (0 to 4) to the second group is 5:1 but the total number of samples is equal to that in the *small* partition. Table 6.8 shows the results of classifiers trained on this partition using no sampling strategy, oversampling (*o*) or sample weighting (*w*).

Compared to the *small* partition, the AUROC of the *10+10* model drops from 0.9907 to 0.9890 and that of the *bin+bin* model from 0.9860 to 0.9845. Sample weighting significantly decreases the performance of the *bin* and *10* models. Oversampling slightly improves the performance of the *bin+10* model but slightly decreases that of the *bin* and *bin+bin* models. The confidence intervals are smaller for training using oversampling hinting at a higher stability. However, more bootstrap iterations are required to rank the two methods with sufficient confidence.

The best-performing model remains *10+10*, trained without oversampling or weighting. Weighting degrades performance but further experiments are required to determine the effect of oversampling for this partition.

6.2.4.3. Between- and within-group imbalance

The *02_13* partition has the same between-group imbalance and the same number of samples as the *02* partition but an additional imbalance in the number of samples of

classifier	sampling	AP	AUROC
10		0.9872 [0.9857, 0.9885]	0.9883 [0.9872, 0.9893]
10	ws	0.9860 [0.9853, 0.9866]	0.9873 [0.9866, 0.9879]
10+10		0.9881 [0.9874, 0.9887]	0.9890 [0.9882, 0.9898]
10+bin		0.9870 [0.9858, 0.9881]	0.9886 [0.9873, 0.9896]
bin+10		0.9843 [0.9794, 0.9881]	0.9860 [0.9816, 0.9894]
bin+10	o	0.9848 [0.9818, 0.9873]	0.9862 [0.9834, 0.9885]
bin		0.9812 [0.9760, 0.9853]	0.9839 [0.9799, 0.9871]
bin	o	0.9804 [0.9762, 0.9839]	0.9831 [0.9801, 0.9856]
bin	w	0.9760 [0.9647, 0.9837]	0.9794 [0.9694, 0.9861]
bin+bin		**0.9817** [0.9768, 0.9856]	**0.9845** [0.9809, 0.9875]
bin+bin	o	0.9811 [0.9772, 0.9844]	0.9842 [0.9816, 0.9864]

Table 6.8.: Results for all permutations of classifier models, permutations and sampling methods trained on the *02* partition, containing 60% of the data with between-class imbalance of 5:1. See table 6.4 for classifier and sampling notation.

different classes in the minority group. Also in this partition, best-performing models were trained on the 10-class problem: *10* and *10+10*. Compared to training on the *02* partition, the AUROC is reduced from 0.9890 to 0.9876 for the *10+10* model and from 0.9845 to 0.9807 for the *bin+bin* model. Note that the AUROC credible intervals widen from [0.9882, 0.9898] to [0.9818, 0.9916] when comparing the *02* and *02_13* partitions for the *10+10* models and from [0.9809, 0.9875] to [0.9737, 0.9859] for the *bin+bin* models, indicating worse average performance and decreased stability when trained on the *02_13* partitions.

Oversampling the second group to account for between-group imbalance improves the performance of the *bin*, *bin+bin*, and *bin+10* models (see table 6.9). The AUROC can be further improved for all 3 models by oversampling that accounts for the within-group imbalance (*os*). Cost weighting decreases the performance of the models *bin* and *10*.

6.2.4.4. Multi-modal model with auxiliary input

All models discussed so far had only access to the images. However, in practice, it might be possible to use additional information, in my example, the class parity. This does not provide the result to the task, whether a sample belongs to the group of categories 0 to 4 or to that of 5 to 9 but presumably makes predicting the group easier. The models *(aux&vis)bin* had access to the class parity during training and require it for testing. These models might learn to rely on this information and could therefore learn an image feature extraction that performs worse than that of models that had only access to the images.

The AUROC of the *(aux&vis)bin* model is 0.9829, which is higher than that of the *bin* model, which achieves 0.9795 (see table 6.9). Hence, the parity carries valuable information either for learning a better model, predicting the groups, or for both. To judge the quality of the lower layers of the vision architecture, it is instructive to assess the performance of the model that was stripped of its auxiliary and final layers, which were replaced and trained in the same way as the *10+bin* and *bin+bin* models. Note

classifier	sampling	AP	AUROC
10		0.9858 [0.9829, 0.9882]	0.9870 [0.9846, 0.9891]
10+10		0.9865 [0.9799, 0.9910]	0.9876 [0.9818, 0.9916]
10+bin		0.9845 [0.9812, 0.9872]	0.9864 [0.9833, 0.9889]
bin+10		0.9807 [0.9697, 0.9878]	0.9828 [0.9735, 0.9889]
bin+10	o	0.9826 [0.9769, 0.9869]	0.9844 [0.9792, 0.9883]
bin+10	os	0.9832 [0.9812, 0.9849]	0.9851 [0.9834, 0.9866]
bin	os	0.9781 [0.9738, 0.9818]	0.9825 [0.9790, 0.9854]
bin+bin	os	0.9790 [0.9768, 0.9810]	0.9829 [0.9811, 0.9845]
(aux&vis)bin		0.9809 [0.9749, 0.9855]	0.9829 [0.9779, 0.9868]
(aux&vis)bin	os	0.9825 [0.9780, 0.9860]	0.9849 [0.9811, 0.9880]
(aux&vis)bin+bin		0.9787 [0.9689, 0.9854]	0.9820 [0.9739, 0.9877]
bin		0.9755 [0.9659, 0.9825]	0.9795 [0.9712, 0.9855]
bin	o	**0.9783** [0.9756, 0.9806]	**0.9821** [0.9789, 0.9848]
bin	w	0.9717 [0.9626, 0.9786]	0.9772 [0.9695, 0.9830]
bin+bin		0.9768 [0.9690, 0.9826]	0.9807 [0.9737, 0.9859]
bin+bin	o	0.9768 [0.9707, 0.9817]	0.9813 [0.9758, 0.9856]

Table 6.9.: Results on the *02_13* partition. Note that *(aux&vis)bin* models performance results can not be compared directly with the other model's performance as they have additional information about the test data. The *(aux&vis)bin+bin* models are tested without auxiliary information. See table 6.4 for classifier and sampling notation.

that this model (*(aux&vis)bin+bin*) does not require the parity of the test images at test time. Freezing the lower layers limited this model to learn to combine the previously learned vision features without learning new features of the slightly different task. This model performs worse than the *10+bin* model but better than the *bin+bin* model. Consequently, auxiliary information helped the network to learn a better vision network but is outperformed by the model that learned to distinguish all 10 categories.

6.2.5. Conclusion

The performance of the classifier degrades with decreasing training data size and within-group and between-group imbalance further limit performance. Cost weighting based on the frequency of class labels is counter productive but oversampling the minority class does improve performance in most cases. Retraining the last layers from scratch is beneficial for model performance in all cases. This is reminiscent of the 2-stage wiring process of neurons connecting to the cortex in human brain development described in section 2.2.3.

The grouping of the categories into the chosen groups causes an imbalance in terms of the variability of appearance of the articles. This might contribute to the higher performance of models that are trained to predict all 10 labels for the task of assigning group (binary) labels. Hence, forcing the networks to learn the harder task of delineating the within-group structure yields better results. This holds true even on the reduced size and imbalanced samples.

An alternative to classifying more categories than needed for the final classification, is

to provide additional information in the form of branches in the network. There is scope for improving medical computer vision models by incorporating additional information such as gender, age or clinical risk factors in the training process. The learned vision architecture can be superior, even in the absence of this information at test-time. In other words, the auxiliary information was useful for the process of learning from the images, not just for the final prediction. Auxiliary information acts as an indirect form of data augmentation and can be seen as the reverse method of multi task learning, which also can improve model performance [Ruder, 2017]. To the best of my knowledge, this has not been reported elsewhere.

6.3. Motion artefact detection - Methods

6.3.1. Diffusion data and annotations

The motion artefact detection algorithm is developed for the application on neonatal diffusion data acquired as part of the Developing Human Connectome Project (dHCP). Subjects are scanned without sedation in natural sleep on a 3T Philips Achieva MR scanner with dedicated 32-channel neonatal head coil and patient handling system [Hughes et al., 2017a]. However, neonates moved in a significant fraction of scans [Hutter et al., 2017].

The multi-shell High Angular Resolution Diffusion Imaging (HARDI) sequence is adapted for the tissue properties of the neonatal brain [Tournier et al., 2015b; Tournier et al., 2015a] and consists of 300 volumes acquired using 4 diffusion weightings: 20 b=0s/mm^2, 64 b=400s/mm^2 88 b=1000s/mm^2 and 128 volumes on the b=2600s/mm^2 shell[4]. The Stejskal-Tanner sequence ($\Delta = 42.5ms, \delta = 14ms, G_{max} = 70mT/m$) uses 4 phase encoding directions and volumes are ordered to maximise the uniformity of the sampling for increased robustness to periodic motion and scan interruption [Hutter et al., 2017]. Volumes have a resolution of $1.5 \times 1.5 \times 3mm^3$, covering a field of view of $15 \times 15 \times 10.2cm^3$ with a slice overlap of 1.5mm (64 slices). The sequence uses multiband factor 4, SENSE 1.2 and partial Fourier 0.855 for acceleration, has an echo spacing of $0.81ms$, and a slice acquisition pattern of interleave 3, shift 2 (see fig. 6.5) [Hutter et al., 2017]. The reconstruction method follows the extended SENSE technique proposed in [Zhu et al., 2016]. Sensitivities were estimated from non-accelerated reference acquisitions with matched readouts as in [Hennel et al., 2016] to promote equivalent distortions in the coil maps as in the data.

For 47 randomly selected neonatal subjects, each volume was annotated manually by displaying the volumes as 2D images in axial, sagittal and coronal projections, in which the point of view can easily be changed, providing more information in case of difficult to assess image quality. Each volume was labelled with respect to motion artefacts as either acceptable, borderline, or unusable ('reject'). No formal rules were defined for assignment to the categories but volumes with multiple slices affected by dropout or severe misalignment between adjacent slices were always rejected, less severe artefacted volumes

[4]For brevity, the units s/mm^2 are dropped in the neural network analysis.

or volumes that are difficult to assess due to high diffusion weighting located between clearly motion artefacted volumes were labelled borderline. Volumes were kept in temporal order but could be traversed in any order for comparison. By labelling volumes in temporal or reverse order, surrounding higher b-value images help in annotating volumes of the b=2600 shell in the case of motion occurring across several consecutive volumes. See fig. 6.6 for annotated coronal images in order of acquisition of an exemplary subject with short bursts and prolonged motion.

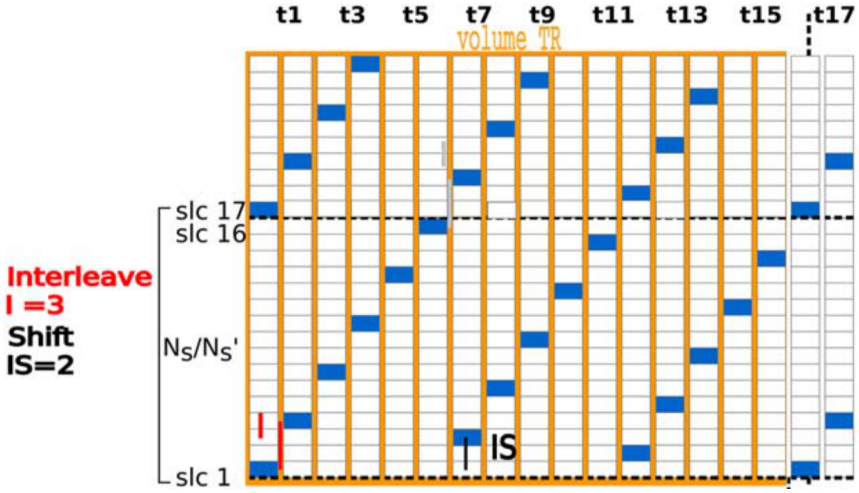


Figure 6.5.: "Illustration of the slice spacing and multiband acquisition order. The slice direction is shown vertically, the excitation order horizontally." [Hutter et al., 2017] Adapted from [Hutter et al., 2017] (Creative Commons Attribution NonCommercial License).

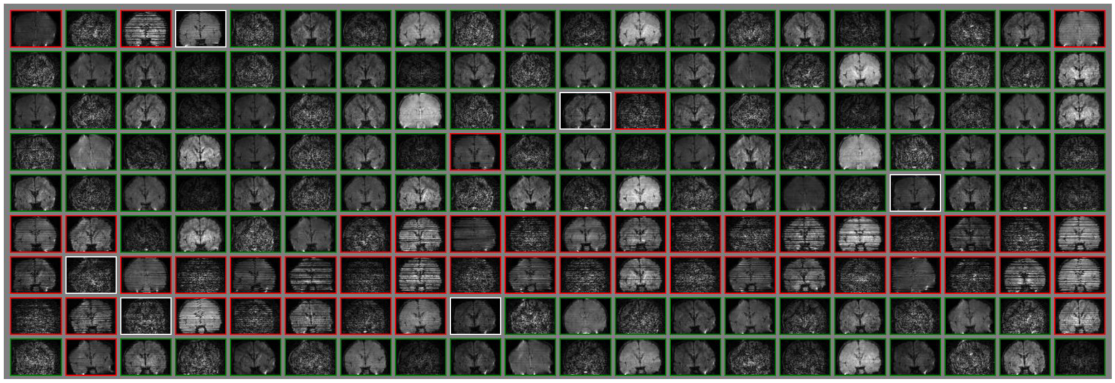


Figure 6.6.: A single coronal slice of 180 consecutive volumes of an exemplary dataset shown in temporal order (row major order) with annotations as colour outline: red ('reject'), white ('borderline'), and green ('accept').

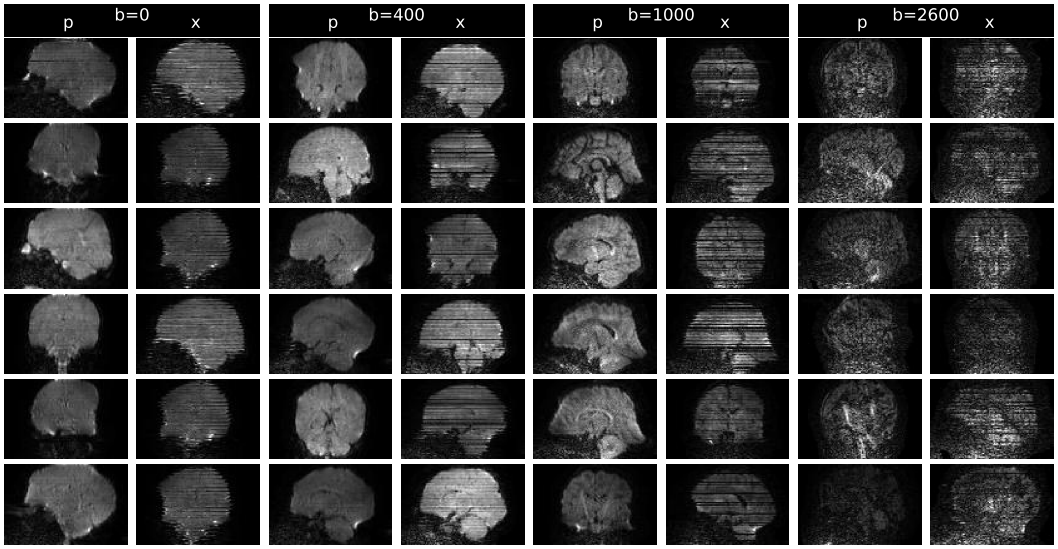


Figure 6.7.: A random sample of the test data. 'p' stands for samples from good volumes (pick) and 'x' for rejected volumes.

Training and validation data split The training and validation set consists of 36 randomly chosen subjects that have on average 274.1, totalling more than 9800 volumes that are labelled either 'accept' or 'reject'. Volumes where the human rater was undecided ('borderline' cases) were not included in training, validation and testing unless otherwise mentioned. The fraction of accept labels among the two clear-cut categories is 85%, 84%, 85% and 90% for the b=0, 400, 1000 and 2600 shell, respectively. Less volumes were labelled as rejected in the highest shell, which is likely a problem of detectability in these noisier data.

For validation, 2 subjects were held out from the training pool and used to assess generalisation quality (learning) and training state selection. These subjects were selected by ranking all subjects' class imbalance and choosing from the subjects with the least overall class imbalance to ensure that the relatively low number of validation subjects contains sufficient examples of each category. Models training on 32 or fewer subjects were monitored using 4 validation subjects.

Each volume was processed and stored as 2D image arrays in coronal and sagittal view through the centre of mass of the original unmasked image as these likely are most representative of the artefacts. The minimum and maximum image intensities of each slice were independently linearly transformed to the range $[0, 1]$ to facilitate learning b-value independent features. For each orientation, an additional six parallel planes adjacent to the centre of mass were added to the data pool, each slice 2 voxels apart from its closest neighbouring slices. This approach aims at increasing the amount of data, improving the robustness of the classification to the location of the centre of mass, and the effect of brain size.

Treating the classification as a 2D problem is motivated by the fact that motion arte-

facts are apparent in 2D projections and do not necessarily require 3D information and can therefore be implemented using much faster 2D convolution filters. The problem could potentially even be treated by 1D convolution filters in the axial direction and pooling in the orthogonal direction. Using multiple slices in each volume remedies the loss of information going from 3D to 2D. Also, pre-trained vision architectures allow transfer learning approaches. To my knowledge, there is no 3D pre-trained network available that is comparable in performance and resource requirements to the popular and high-performing networks pre-trained on 2D RGB general purpose images.

Using multiple sagittal and coronal projects results in more than 138 000 images with relatively high variation of location and size of anatomical features. However, all slices from a single volume are treated as multiple examples of the same data source and slices of the same subject are assigned to one of training, validation or testing data pools as spreading images across set boundaries might prevent detection of overfitting of subject-specific markers such as specific anatomical features. All training images were randomly shuffled before each training epoch and the model did not have access to information such as slice location or subject data other than the bare image.

Test data The test data consists of 3201 volumes from 11 randomly selected subjects. Of the 210, 684, 947 and 1360 volumes in each shell in order of increasing b-value, 25, 93, 132 and 118 were labelled as rejected. The respective fraction of accepted volumes is 88%, 86%, 86% and 91%.

To test human operator variability, 4 subjects were labelled again by the same operator (me) and additionally by a second operator (Christopher Kelly, MD, PhD) who has annotated a large number of volumes himself and seen a representative sample of classifications of non-test subjects annotated by the first operator. The inter-operator variability dataset contains 1072 volumes labelled as 'accept' or 'reject' and the repeat annotations contain 1154 volumes not labelled as 'borderline'.

6.3.2. Training and testing setup

All networks were trained on images with intensities scaled between 0 and 1 for each slice independently and unless otherwise stated, training data was augmented using random geometric transformations: horizontal shifting by up to 10% of the FOV, horizontal flipping and zooming between 90% and 110% of the original image size. Missing areas are filled by reflecting the augmented image to avoid artificial sharp boundaries in the image (see fig. 6.8). In contrast to the original VGG16 network, the Keras version of the model uses equally scaled colour channels. Therefore, RGB input for the VGG16 networks was simulated by concatenating the grey-scale channel 3 times. Test image augmentation uses 5% shifting and zooming, which simulates the image characteristics of augmented training images while being in the centre of the training distortion space.

Except for the fixed parameters in the transfer learning models, all bias terms of all models are initialised with zeros and all weights are initialised using Glorot (a.k.a Xavier) initialisation [Glorot, Bengio, 2010]. Glorot initialisation draws weights from a uniform

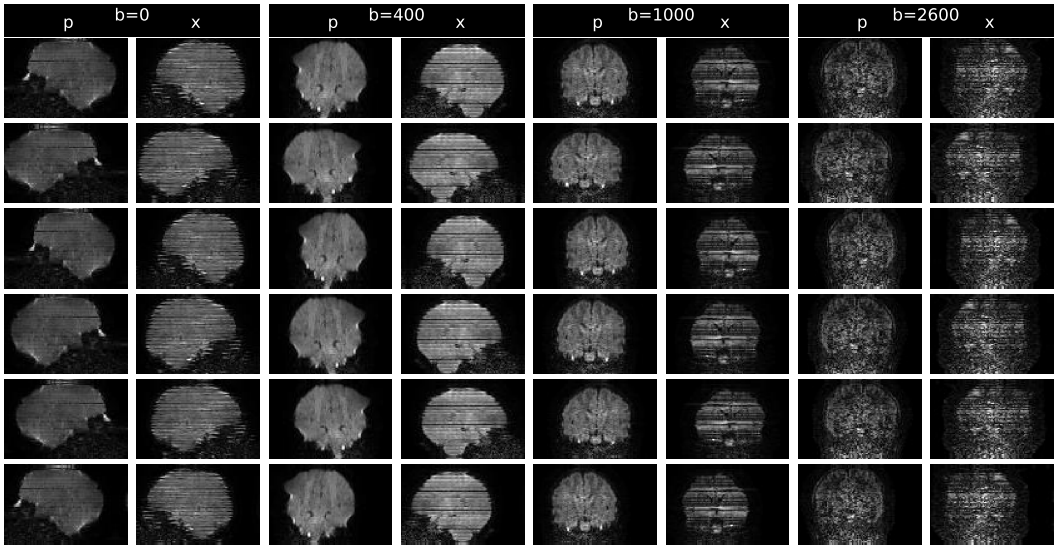


Figure 6.8.: A sample of the non-augmented training data of each b-value and class (top row) and multiple augmented versions of the same images below. 'p' stands for good volumes (pick) and 'x' for rejected volumes.

distribution with limits $\pm\sqrt{6/(\text{number of input units} + \text{number of output units})}$, ensuring that the variance of the gradients does not vanish or increase exponentially as they pass through the layers during training [He et al., 2015b]. This reduces model volatility especially for deep neural networks but does not completely prevent models from becoming corrupted during training.

Unless otherwise stated, all models were trained with batches containing an equal number of good and outlier volumes and an equal number of images across b-values. Training of all models was performed in batches of 64 images, which is a typical batch size [Goodfellow, Bengio, Courville, 2016, chapter 8] and motivated by my GPU memory capacity and processing speed when training the largest VGG16 network. Batch size was reduced to 60 for models trained on 3 of the 4 b-values to allow a balanced number of images in each batch. An epoch ends when all data has been presented to the network at least once.

All models were trained using a cross-entropy loss metric optimised by the Adam algorithm [Kingma, Ba, 2014] with exponential first and second order moment decay parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, without learning rate decay and with a learning rate of 0.001 for all models trained from scratch except for the high parametric models (*scratch_2233d*, *scratch_22333d*, see section 6.3.3.2), for which the learning rate was reduced by a factor of 10 to increase training robustness. Transfer learning and fine tuning of the pre-trained VGG16 network was performed with slow learning rate of 0.0001, except for models that reuse the early layers of the VGG16 model (for instance *vgg16_2*), for which training was found to be more robust with a learning rate of 0.0005. Convergence in training loss was reached for most models before 30 iterations but training was continued until at least 30 epochs were reached. Learning rate and batch size in-

fluence the width and depth of training cost function minima found by neural networks trained using batched gradient descent [Jastrzębski et al., 2017]. However, it is not clear whether or how that affects generalisation performance [Dinh et al., 2017]. Also, Adam adjusts the parameter update dynamically based on the first and second order moments of the gradient and is relatively robust to the choice of parameters [Kingma, Ba, 2014]. Optimising learning rate, batch size and learning rate decay is left for future work.

Additionally to the final model state, learned parameters were saved after 10, 20 and 30 epochs and if any of AUROC, AP, or accuracy on the validation set increased between epochs. Training was continued beyond the minimum number of epochs until no validation metric decreased for 10 consecutive epochs. Training beyond convergence is motivated by work showing networks tend to learn information compression in this phase, which improves generalisation performance [Mehta, Schwab, 2014; Shwartz-Ziv, Tishby, 2017].

For all models, training was repeated 3 times with resampled training and validation data. To reduce sample selection variation for model comparison, all models were trained and evaluated on the same disjoint sets of data for the respective training repetition.

For all models, the state that achieved the highest AUROC on the validation set was used for performance evaluation on the test set. However, models with a sudden increase in training loss (possibly due to vanishing or exploding gradients) or with AUROC on the validation set of less than 0.8 were excluded from the analysis. Failed training runs were not repeated. Unless stated otherwise, all results reported are derived from the 3 training repetitions.

The default parameters reported in this section were selected by experimentation with training stability across all model architectures. However, all experiments were designed before any model was evaluated on the test set to prevent bias in the parameter search. During the experiment design and training, the only information available about generalisation performance was the training and validation loss on 2 or 4 subjects. Previous work on the same dataset [Kelly et al., 2017] involved a different network architecture and, except for the slice-selection, explored different aspects related to transfer-learning and architecture fusion.

6.3.3. Model architecture search space

Similar to the class-imbalance experiments, the motion artefact detection is performed with VGG16-like (see section 6.1.3 and table 6.1) or at least inspired by its design. The CNN architectures investigated can be grouped by their design principles and source of training data into 3 categories.

6.3.3.1. Models derived from pre-trained VGG16 network

The original version of the VGG16 network (table 6.1) was designed to classify colour photos of the imagenet dataset into 1000 categories. Training this network on that dataset took 2-3 weeks on 4 high performing GPUs [Simonyan, Zisserman, 2014]. I used the weights of this VGG16 network to determine whether it is possible to transfer

the learned feature representation and extraction to the motion artefact classification problem.

Since MRI images are not part of the imagenet classes, the last layer has to be retrained to differentiate between 2 instead of 1000 classes. The VGG16 network adjusted to 2 output channels and 64 by 99 pixel input is shown in table 6.10. In order to feed greyscale images into the network, individual slices were stacked 3 times at their native resolution. However, the original network expects an input size of 224 by 224 pixels, which is higher than the resolution of the diffusion images. Since convolution layers connect features on a per-pixel basis, it is possible to reuse all convolution layer weights as they are. The only limitation is a minimum input size due to the spatial 2x2 pooling layers with stride 2 that reduce the spatial dimension after each block. The last convolution block of the modified network has a spatial extent of 4 by 6 instead of 14 by 14 which leads to a 2 by 3 image after pooling instead of the 7 by 7 feature map. Consequently, the layers following the flattening layer after the last convolution block have to be adjusted in their dimensionality (after line 20 in table 6.1) and retrained from scratch. I also reduced the number of units in the final fully connected layers from 4096 to 16 to reduce the number of parameters and account for the presumably lower requirements in feature combination complexity of a binary classification task than the original classification into 1000 categories. This approach allows reusing all of the feature extraction engine in the first 13 layers and training a comparably low-parametric classifier on the 25088 features produced by the network up to layer 20. The networks that reuse all the pre-trained convolution layer weights but retrain only the 3 final classification layers with the reduced number of units are denoted as *vgg16_22333*. Note that an alternative approach would be to upsample and crop the diffusion images to the network's native resolution but this approach might degrade image textures and was not explored.

Table 6.10.: The VGG16 network architecture, adjusted for 2 classes and with 16 instead of 4096 units in the fully connected dense layers. (Compare to the last layers of the original VGG16 network shown in table 6.1.)

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 3)	0
2	Conv2D 3x3	relu	(64, 99, 64)	1792
3	Conv2D 3x3	relu	(64, 99, 64)	36928
4	MaxPooling2D 2x2		(32, 49, 64)	0
5	Conv2D 3x3	relu	(32, 49, 128)	73856
6	Conv2D 3x3	relu	(32, 49, 128)	147584
7	MaxPooling2D 2x2		(16, 24, 128)	0
8	Conv2D 3x3	relu	(16, 24, 256)	295168
9	Conv2D 3x3	relu	(16, 24, 256)	590080
10	Conv2D 3x3	relu	(16, 24, 256)	590080
11	MaxPooling2D 2x2		(8, 12, 256)	0
12	Conv2D 3x3	relu	(8, 12, 512)	1180160
13	Conv2D 3x3	relu	(8, 12, 512)	2359808
14	Conv2D 3x3	relu	(8, 12, 512)	2359808
15	MaxPooling2D 2x2		(4, 6, 512)	0
16	Conv2D 3x3	relu	(4, 6, 512)	2359808
17	Conv2D 3x3	relu	(4, 6, 512)	2359808
18	Conv2D 3x3	relu	(4, 6, 512)	2359808
19	MaxPooling2D 2x2		(2, 3, 512)	0
20	Flatten		(3072,)	0
21	Dense	linear	(16,)	49168
22	Dense	relu	(16,)	272
23	Dropout (p=0.5)		(16,)	0
24	Dense	sigmoid	(2,)	34

Motion artefact classification presumably requires less hierarchical feature representations and smaller spatial context than object detection. Therefore, it might be sufficient to reuse only low-level features produced by the first blocks and train a classifier on their output. However, the original VGG16 network is designed to reduce the spatial extent of the feature maps gradually between blocks. Consequently, reusing only the first layers poses the design question of whether the spatial domain is to be reduced by pooling in a single step, potentially loosing significant amount of information, or to be fed as a 1D vector to the final classifier layers, loosing no information other than the spatial context but increasing the size of the final classification layers. Two network architectures using the output of the second block for classification are shown in table 6.12. The former architecture spatially averages the 16x24 feature map before it is flattened, while the latter concatenates the feature vectors of all pixels. Concatenation increases the number of parameters in the first dense layer from 2064 to 786448. Global average pooling could be replaced by maximum pooling or by a local pooling that retains some of the spatial information. Exploring these options is left for future work. Networks reusing the unchanged weights of the first 1, 2, 3, 4 and 5 blocks with global average pooling are named *vgg16_2*, *vgg16_22*, *vgg16_223*, *vgg16_2233*, and *vgg16_22333*, indicating the number of convolution layers in each block. Networks that do not use global average pooling have the suffix *nopool*. In the networks *vgg16_2233f3* and *vgg16_f22333*, all layers after the letter *f* were fine-tuned using the VGG16 weights as initialisation, after training the final 3 dense layers.

Table 6.11.: The *vgg16_2* network architecture, reusing the first block of the VGG16 network followed by global average pooling. None of the attempts to train this network without global average pooling were successful.

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 3)	0
2	Conv2D 3x3	relu	(64, 99, 64)	1792 (1792)
3	Conv2D 3x3	relu	(64, 99, 64)	36928 (36928)
4	MaxPooling2D 2x2		(32, 49, 64)	0
5	Sequential		(2,)	1346

Table 6.12.: *vgg16_22* with (left) and without (right) global average pooling. The top part of the table is identical for both networks and kept fixed during training (indicated by the number of non-trainable parameters in brackets).

	name	activation	output shape	parameters	output shape	parameters
1	InputLayer		(64, 99, 3)	0	(64, 99, 3)	0
2	Conv2D 3x3	relu	(64, 99, 64)	1792 (1792)	(64, 99, 64)	1792 (1792)
3	Conv2D 3x3	relu	(64, 99, 64)	36928 (36928)	(64, 99, 64)	36928 (36928)
4	MaxPooling2D 2x2		(32, 49, 64)	0	(32, 49, 64)	0
5	Conv2D 3x3	relu	(32, 49, 128)	73856 (73856)	(32, 49, 128)	73856 (73856)
6	Conv2D 3x3	relu	(32, 49, 128)	147584 (147584)	(32, 49, 128)	147584 (147584)
7	MaxPooling2D 2x2		(16, 24, 128)	0	(16, 24, 128)	0
8	AveragePooling2D 15x23		(1, 1, 128)	0	-	-
9	Flatten		(128,)	0	(49152,)	0
10	Dense	linear	(16,)	2064	(16,)	786448
11	Dense	relu	(16,)	272	(16,)	272
12	Dropout (p=0.5)		(16,)	0	(16,)	0
13]	Dense	sigmoid	(2,)	34	(2,)	34

Table 6.13.: *vgg16_223* with (left) and without (right) global pooling.

	name	activation	output shape	parameters	output shape	parameters
1	InputLayer		(64, 99, 3)	0	(64, 99, 3)	0
2	Conv2D 3x3	relu	(64, 99, 64)	1792 (1792)	(64, 99, 64)	1792 (1792)
3	Conv2D 3x3	relu	(64, 99, 64)	36928 (36928)	(64, 99, 64)	36928 (36928)
4	MaxPooling2D 2x2		(32, 49, 64)	0	(32, 49, 64)	0
5	Conv2D 3x3	relu	(32, 49, 128)	73856 (73856)	(32, 49, 128)	73856 (73856)
6	Conv2D 3x3	relu	(32, 49, 128)	147584 (147584)	(32, 49, 128)	147584 (147584)
7	MaxPooling2D 2x2		(16, 24, 128)	0	(16, 24, 128)	0
8	Conv2D 3x3	relu	(16, 24, 256)	295168 (295168)	(16, 24, 256)	295168 (295168)
9	Conv2D 3x3	relu	(16, 24, 256)	590080 (590080)	(16, 24, 256)	590080 (590080)
10	Conv2D 3x3	relu	(16, 24, 256)	590080 (590080)	(16, 24, 256)	590080 (590080)
11	MaxPooling2D 2x2		(8, 12, 256)	0	(8, 12, 256)	0
12	AveragePooling2D 7x11		(1, 1, 256)	0	-	-
13	Flatten		(256,)	0	(24576,)	0
14	Dense	linear	(16,)	4112	(16,)	393232
15	Dense	relu	(16,)	272	(16,)	272
16	Dropout (p=0.5)		(16,)	0	(16,)	0
17	Dense	sigmoid	(2,)	34	(2,)	34

Table 6.14.: *vgg16_2233* with (left) and without (right) global pooling.

	name	activation	output shape	parameters	output shape	parameters
1	InputLayer		(64, 99, 3)	0	(64, 99, 3)	0
2	Conv2D 3x3	relu	(64, 99, 64)	1792 (1792)	(64, 99, 64)	1792 (1792)
3	Conv2D 3x3	relu	(64, 99, 64)	36928 (36928)	(64, 99, 64)	36928 (36928)
4	MaxPooling2D 2x2		(32, 49, 64)	0	(32, 49, 64)	0
5	Conv2D 3x3	relu	(32, 49, 128)	73856 (73856)	(32, 49, 128)	73856 (73856)
6	Conv2D 3x3	relu	(32, 49, 128)	147584 (147584)	(32, 49, 128)	147584 (147584)
7	MaxPooling2D 2x2		(16, 24, 128)	0	(16, 24, 128)	0
8	Conv2D 3x3	relu	(16, 24, 256)	295168 (295168)	(16, 24, 256)	295168 (295168)
9	Conv2D 3x3	relu	(16, 24, 256)	590080 (590080)	(16, 24, 256)	590080 (590080)
10	Conv2D 3x3	relu	(16, 24, 256)	590080 (590080)	(16, 24, 256)	590080 (590080)
11	MaxPooling2D 2x2		(8, 12, 256)	0	(8, 12, 256)	0
12	Conv2D 3x3	relu	(8, 12, 512)	1180160 (1180160)	(8, 12, 512)	1180160 (1180160)
13	Conv2D 3x3	relu	(8, 12, 512)	2359808 (2359808)	(8, 12, 512)	2359808 (2359808)
14	Conv2D 3x3	relu	(8, 12, 512)	2359808 (2359808)	(8, 12, 512)	2359808 (2359808)
15	MaxPooling2D 2x2		(4, 6, 512)	0	(4, 6, 512)	0
16	AveragePooling2D 3x5		(1, 1, 512)	0	-	-
17	Flatten		(512,)	0	(12288,)	0
18	Dense	linear	(16,)	8208	(16,)	196624
19	Dense	relu	(16,)	272	(16,)	272
20	Dropout (p=0.5)		(16,)	0	(16,)	0
21	Dense	sigmoid	(2,)	34	(2,)	34

6.3.3.2. The VGG architectures trained from scratch

Models named *scratch_** are heavily inspired by the design of the VGG16 network. The main difference is that they have between 5 and 16 fully connected layers and only 8 instead of 64 units in layers of the first convolution block. Similar to the VGG16 network, this number is doubled in each following block. The last block is followed by a flattening layer and 3 fully connected layers, which all use non-linear activation functions. They are named scratch because all models were trained from scratch and the numbers after the underscore indicate the number of convolution layers in each block. The output of the last local maximum pooling is fed as a 1D vector into the final dense layers without additional spatial pooling. Consequently, networks with lower number of blocks have larger fully connected layers. The extreme case is the model that consists of only one block (*scratch_2*), which has 100360 of its 101114 parameters in the first dense layer. In contrast, the model consisting of 5 convolution blocks (*scratch_22333d*) has only 6152 parameters in the first dense layer. The models *scratch_223* and *scratch_2233* are the most similar in their size of the final convolution layer's spatial map to the original VGG16 network. See tables 6.15, 6.31 to 6.33, 6.35 and 6.36 for details of the model architectures.

In order to prevent overfitting of the deeper networks, the models with 4 or more blocks were trained using dropout regularisation [Srivastava et al., 2014] on the output of some or of all maximum pooling layers and before the final classification layer. The dropout rate (p) was set between 0.2 and 0.5 as this gave a smooth decrease in training loss but neither the locations, nor the magnitude of the regularisation were optimised. The four block architecture was implemented without (*scratch_2233*, table 6.15) and

with dropout regularisation (*scratch_2233d*, table 6.35).

Additionally a four block network with 32 input units was trained using dropout regularisation. This architecture (*scratch_2233d32*, table 6.37) has four times the number of units in the first layer as the other *scratch* networks but half the number as that of the VGG16 network.

Table 6.15.: The *scratch_2233* model architecture.

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 1)	0
2	Conv2D 3x3	relu	(64, 99, 8)	80
3	Conv2D 3x3	relu	(64, 99, 8)	584
4	MaxPooling2D 2x2		(32, 49, 8)	0
5	Conv2D 3x3	relu	(32, 49, 16)	1168
6	Conv2D 3x3	relu	(32, 49, 16)	2320
7	MaxPooling2D 2x2		(16, 24, 16)	0
8	Conv2D 3x3	relu	(16, 24, 32)	4640
9	Conv2D 3x3	relu	(16, 24, 32)	9248
10	Conv2D 3x3	relu	(16, 24, 32)	9248
11	MaxPooling2D 2x2		(8, 12, 32)	0
12	Conv2D 3x3	relu	(8, 12, 64)	18496
13	Conv2D 3x3	relu	(8, 12, 64)	36928
14	Conv2D 3x3	relu	(8, 12, 64)	36928
15	MaxPooling2D 2x2		(4, 6, 64)	0
16	Flatten		(1536,)	0
17	Dense	relu	(8,)	12296
18	Dense	relu	(8,)	72
19	Dense	sigmoid	(2,)	18

6.3.3.3. The *custom*-made architecture

A third class of network architectures deviates from the VGG architecture the most. It was developed following the intuition that a model with a smaller number of layers and more filters in layers closer to the input layer might perform better on the relatively small dataset and the binary classification problem of mostly textural information. The goal is to learn robust representations on a small dataset that distinguish between artefacted and good data and to train a network that does not overfit to a particular spatial arrangements of these patterns present in the training data.

Two types of regularisation are used to ensure stable training and to prevent overfitting: The output of all convolution and all but the last dense layer are fed into batch normalisation layers before applying the non-linearity (rectified linear activation layers). A batch normalisation layer (see section 4.4.1) performs two operations. First, it scales and shifts its input linearly (affine transformation) to zero mean and unit variance across all samples in the batch. Then it applies the learned affine transformation to the normalised data. Hence, it increases the learned parameter count by 2 for each input channel and requires computing and storing the mean and variance of a particular batch, adding another 2 parameters required for training. During testing, the normalisation to zero mean and unit variance is applied to a single sample but could use test data statistics as well [Ioffe, Szegedy, 2015]. This normalisation has the effect of stabilising the learning, leading to faster training convergence and better performing models [Ioffe, Szegedy,

2015]. Through batch-wise normalisation, training samples are always seen in the context of other samples. This adds a stochastic variation of features and can be seen as feature augmentation and is likely an important aspect in the context of class imbalance where oversampling the minority class stabilises the batch sample statistics. The original implementation uses batch normalisation layers after the non-linearity [Ioffe, Szegedy, 2015]. I opted for placing it in between convolution and activation layers, which yielded higher performance in the Inception-v4 model [Szegedy et al., 2017].

Each rectified linear unit layer is followed by a dropout regularisation layer in which between 10% and 20% of the features are set to zero. Note, that recent research suggests that the combination of batch normalisation and dropout can lead to suboptimal test performance [Li et al., 2018]. I did not test the *custom* architecture without dropout regularisation or without batch normalisation.

To investigate the influence of the shape and size of convolution filters, and the effect of pooling, I trained different versions of this network architecture. Models named *custom* and *custom2* use 1D convolution filters in the first convolution layer, which span 5 and 3 pixels in the axial direction followed by 4x1 and 2x1 max pooling layers, respectively. The following convolution and pooling layers use 2x2 filters, increasing the receptive field in the non-axial direction and reducing the spatial width of the feature maps to 23 pixels. Similarly to the *scratch* networks, the last dense layers use 8 units to combine the flattened feature maps. See table 6.16 for a list of the layers in these architectures.

The architectures *custom3* and *custom4* (table 6.17) use square (3x3) and horizontally oriented 1D (1x3) convolution filters in the first layer and 2x2 and 1x2 spatial pooling. Finally, models of the *custom5* type use 3x1 convolution filters and 2x1 pooling layers throughout, preserving the full spatial resolution in the non-axial direction (table 6.18).

Table 6.16.: *custom* (left) and *custom2* (right) architectures.

	name	activation	output shape	parameters	name	output shape	parameters
1	InputLayer		(64, 99, 1)	0	InputLayer	(64, 99, 1)	0
2	Conv2D 5x1	linear	(60, 99, 64)	320	Conv2D 3x1	(62, 99, 64)	192
3	BatchNormalization		(60, 99, 64)	256 (128)	BatchNormalization	(62, 99, 64)	256 (128)
4	Activation	relu	(60, 99, 64)	0	Activation	(62, 99, 64)	0
5	MaxPooling2D 4x1		(15, 99, 64)	0	MaxPooling2D 2x1	(31, 99, 64)	0
6	Dropout (p=0.1)		(15, 99, 64)	0	Dropout (p=0.1)	(31, 99, 64)	0
7	Conv2D 3x3	linear	(13, 97, 48)	27648	Conv2D 3x3	(29, 97, 48)	27648
8	BatchNormalization		(13, 97, 48)	192 (96)	BatchNormalization	(29, 97, 48)	192 (96)
9	Activation	relu	(13, 97, 48)	0	Activation	(29, 97, 48)	0
10	MaxPooling2D 2x2		(6, 48, 48)	0	MaxPooling2D 2x2	(14, 48, 48)	0
11	Dropout (p=0.1)		(6, 48, 48)	0	Dropout (p=0.1)	(14, 48, 48)	0
12	Conv2D 3x3	linear	(4, 46, 32)	13824	Conv2D 3x3	(12, 46, 32)	13824
13	BatchNormalization		(4, 46, 32)	128 (64)	BatchNormalization	(12, 46, 32)	128 (64)
14	Activation	relu	(4, 46, 32)	0	Activation	(12, 46, 32)	0
15	MaxPooling2D 2x2		(2, 23, 32)	0	MaxPooling2D 2x2	(6, 23, 32)	0
16	Dropout (p=0.1)		(2, 23, 32)	0	Dropout (p=0.1)	(6, 23, 32)	0
17	Flatten		(1472,)	0	Flatten	(4416,)	0
18	Dense	linear	(8,)	11776	Dense	(8,)	35328
19	BatchNormalization		(8,)	32 (16)	BatchNormalization	(8,)	32 (16)
20	Activation	relu	(8,)	0	Activation	(8,)	0
21	Dropout (p=0.2)		(8,)	0	Dropout (p=0.2)	(8,)	0
22	Dense	linear	(8,)	64	Dense	(8,)	64
23	BatchNormalization		(8,)	32 (16)	BatchNormalization	(8,)	32 (16)
24	Activation	relu	(8,)	0	Activation	(8,)	0
25	Dropout (p=0.2)		(8,)	0	Dropout (p=0.2)	(8,)	0
26	Dense	linear	(2,)	16	Dense	(2,)	16
27	Activation	sigmoid	(2,)	0	Activation	(2,)	0

Table 6.17.: *custom3* (left) and *custom4* (right) architectures.

	name	activation	output shape	parameters	name	output shape	parameters
1	InputLayer		(64, 99, 1)	0	InputLayer	(64, 99, 1)	0
2	Conv2D 3x3	linear	(62, 97, 64)	576	Conv2D 1x3	(64, 97, 64)	192
3	BatchNormalization		(62, 97, 64)	256 (128)	BatchNormalization	(64, 97, 64)	256 (128)
4	Activation	relu	(62, 97, 64)	0	Activation	(64, 97, 64)	0
5	MaxPooling2D 2x2		(31, 48, 64)	0	MaxPooling2D 1x2	(64, 48, 64)	0
6	Dropout (p=0.1)		(31, 48, 64)	0	Dropout (p=0.1)	(64, 48, 64)	0
7	Conv2D 3x3	linear	(29, 46, 48)	27648	Conv2D 3x3	(62, 46, 48)	27648
8	BatchNormalization		(29, 46, 48)	192 (96)	BatchNormalization	(62, 46, 48)	192 (96)
9	Activation	relu	(29, 46, 48)	0	Activation	(62, 46, 48)	0
10	MaxPooling2D 2x2		(14, 23, 48)	0	MaxPooling2D 2x2	(31, 23, 48)	0
11	Dropout (p=0.1)		(14, 23, 48)	0	Dropout (p=0.1)	(31, 23, 48)	0
12	Conv2D 3x3	linear	(12, 21, 32)	13824	Conv2D 3x3	(29, 21, 32)	13824
13	BatchNormalization		(12, 21, 32)	128 (64)	BatchNormalization	(29, 21, 32)	128 (64)
14	Activation	relu	(12, 21, 32)	0	Activation	(29, 21, 32)	0
15	MaxPooling2D 2x2		(6, 10, 32)	0	MaxPooling2D 2x2	(14, 10, 32)	0
16	Dropout (p=0.1)		(6, 10, 32)	0	Dropout (p=0.1)	(14, 10, 32)	0
17	Flatten		(1920,)	0	Flatten	(4480,)	0
18	Dense	linear	(8,)	15360	Dense	(8,)	35840
19	BatchNormalization		(8,)	32 (16)	BatchNormalization	(8,)	32 (16)
20	Activation	relu	(8,)	0	Activation	(8,)	0
21	Dropout (p=0.2)		(8,)	0	Dropout (p=0.2)	(8,)	0
22	Dense	linear	(8,)	64	Dense	(8,)	64
23	BatchNormalization		(8,)	32 (16)	BatchNormalization	(8,)	32 (16)
24	Activation	relu	(8,)	0	Activation	(8,)	0
25	Dropout (p=0.2)		(8,)	0	Dropout (p=0.2)	(8,)	0
26	Dense	linear	(2,)	16	Dense	(2,)	16
27	Activation	sigmoid	(2,)	0	Activation	(2,)	0

Table 6.18.: *custom5* architecture.

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 1)	0
2	Conv2D 3x1	linear	(62, 99, 64)	192
3	BatchNormalization		(62, 99, 64)	256 (128)
4	Activation	relu	(62, 99, 64)	0
5	MaxPooling2D 2x1		(31, 99, 64)	0
6	Dropout (p=0.1)		(31, 99, 64)	0
7	Conv2D 3x1	linear	(29, 99, 48)	9216
8	BatchNormalization		(29, 99, 48)	192 (96)
9	Activation	relu	(29, 99, 48)	0
10	MaxPooling2D 2x1		(14, 99, 48)	0
11	Dropout (p=0.1)		(14, 99, 48)	0
12	Conv2D 3x1	linear	(12, 99, 32)	4608
13	BatchNormalization		(12, 99, 32)	128 (64)
14	Activation	relu	(12, 99, 32)	0
15	MaxPooling2D 2x1		(6, 99, 32)	0
16	Dropout (p=0.1)		(6, 99, 32)	0
17	Flatten		(19008,)	0
18	Dense	linear	(8,)	152064
19	BatchNormalization		(8,)	32 (16)
20	Activation	relu	(8,)	0
21	Dropout (p=0.2)		(8,)	0
22	Dense	linear	(8,)	64
23	BatchNormalization		(8,)	32 (16)
24	Activation	relu	(8,)	0
25	Dropout (p=0.2)		(8,)	0
26	Dense	linear	(2,)	16
27	Activation	sigmoid	(2,)	0

6.4. Motion artefact detection - Experiments

I perform multiple training experiments on VGG-architecture inspired models, designed to gain insights into whether deep networks and large amount of data are required for this task, whether transfer learning is beneficial for classification performance, and how to best train these networks given the class imbalance in the diffusion data. In particular, the experiments investigate (a) the effect of model depth, (b) the number of model parameters, (c) data sampling and weighting, and (d) image augmentation schemes on the classification performance of diffusion motion artefacts.

6.4.1. Defining model evaluation strategies: Metrics, slice-selection and slice-pooling

The CNN classifiers were trained to assign labels to 2D slices of volumes. However, the final application is to assign labels to volumes. The question at heart is: how to best transition from the training regime that uses 2D, possibly distorted images from multiple locations in the brain, to the test regime that is interested in single volume-level labels?

The decision on whether to reject a volume can be made using, for instance, a subset of the slices or the average classification result across slices, or any other combination of classification results such as median, maximum norm or a decision tree. It seems plausible

that a consensus vote across multiple slices can lead to better and more stable predictions. However, slices close to the centre of the brain might provide better classification results than slices at the periphery because they provide more useful data. During ground truth labelling, human observers saw all 3 projections in the centre of the image and were able to scroll through the images. I tended to look at the sagittal projection first. One might find a changing "preference" between coronal or sagittal slices and the location of the slice varying across classifiers. If a model-independent preference for certain slices exists, it would be valuable to know for future training image sampling schemes.

However, these aspects are also important for establishing an unbiased test of model performance. For instance, imagine a classifier that quantifies the total *amount* of artefacts present in the image and uses a simple threshold to decide whether an image is artefacted or not. This classifier would yield biased results in slices far from the centre of the brain as the potentially artefacted area is small compared to the field of view. Deep neural networks tend to need large amounts of training data which can be achieved by image augmentation which increases the amount of data but also introduces blurring due to resampling. These networks therefore learn a mapping from blurred images to labels and using them on the original non-blurred data might decrease their performance or cause biased results if they are sensitive to this change in image properties. However, using non-augmented data could be beneficial for stability if it preserves the artefacts' characteristics. This effect could be dependent on the network architecture and therefore bias the model selection process.

The weighting of model stability is another consideration that is affected by the choice of testing scheme. Consider a classifier that produces random predictions in 50% of the slices but retrieves the correct label with likelihood 100% in all other slices. It would appear unstable in the evaluation of a single slice but might be more useful if the labels are averaged on the slice and orientation-level compared to a more stable classifier with lower average performance. Also, pooling multiple augmented versions of each image can improve overall classification performance significantly [Valle et al., 2017]. However, if computation time during inference is a bottleneck, then the more stable algorithm might be preferable. This weighting of stability against computing resources is implicit in whether one evaluates on single slices or across pooled decisions.

Definitely answering these questions would require additional testing data to determine which model evaluation techniques lead to better models and is outside the scope of this work. However, it is instructive to investigate how performance metrics for a representative sample of model architectures vary across different testing conditions. This gives some insights into the structure of the data and on the performance metrics.

Table 6.19 shows the test results of a single model (*scratch_22*) split up by b-value and evaluated across all b-values (b=*all*). The classification labels were generated using augmented and non-augmented test data as indicated in the 'aug' column. The classification results can either be compared at the slice level or at the volume level. Volume-level labels were derived from single individual slices in the centre of mass or by averaging the classifications of all 14 coronal and sagittal slices. Additionally, for the test scenarios using augmentations, the analysis was performed on the average classification of 25 differently augmented versions of the test data.

b	aug	sample	H	$H^{\vee}$	AP	AUROC	sp95r	MCC	TN	FN	TP	FP
0	yes	1, com:sag	0.935	0.903	0.974	0.996	0.968	0.836	176	0	25	9
0	no	1, com:sag	0.944	0.9	0.973	0.996	0.978	0.836	176	0	25	9
0	yes	25, com:sag	0.935	0.903	0.974	0.996	0.968	0.836	176	0	25	9
0	yes	1, com:cor	0.897	0.853	0.958	0.993	0.951	0.822	175	0	25	10
0	yes	1, group:s+v	0.908	0.871	0.968	0.995	0.962	0.822	175	0	25	10
0	no	1, group:s+v	0.913	0.877	0.964	0.994	0.973	0.822	175	0	25	10
0	yes	25, com:cor	0.897	0.853	0.958	0.993	0.951	0.822	175	0	25	10
0	yes	25, group:s+v	0.913	0.877	0.969	0.995	0.973	0.822	175	0	25	10
0	no	1, slices	0.886	0.833	0.956	0.993	0.962	0.812	2449	5	345	141
0	yes	1, slices	0.887	0.836	0.957	0.993	0.957	0.805	2438	3	347	152
0	no	1, com:cor	0.891	0.833	0.944	0.991	0.962	0.783	174	1	24	11
400	yes	25, group:s+v	0.993	0.987	0.996	0.999	0.998	0.942	581	0	93	10
400	yes	1, group:s+v	0.997	0.993	0.997	1	0.998	0.937	580	0	93	11
400	yes	1, com:sag	0.993	0.987	0.992	0.999	0.998	0.926	578	0	93	13
400	no	1, group:s+v	0.997	0.993	0.996	0.999	0.998	0.926	578	0	93	13
400	yes	25, com:sag	0.993	0.987	0.992	0.999	0.998	0.926	578	0	93	13
400	no	1, slices	0.986	0.976	0.996	0.999	0.997	0.921	8078	0	1302	196
400	no	1, com:sag	0.984	0.973	0.995	0.999	0.998	0.916	576	0	93	15
400	no	1, com:cor	0.993	0.986	0.996	0.999	0.997	0.916	576	0	93	15
400	yes	1, slices	0.983	0.971	0.995	0.999	0.997	0.912	8053	0	1302	221
400	yes	1, com:cor	0.981	0.97	0.995	0.999	0.997	0.911	575	0	93	16
400	yes	25, com:cor	0.981	0.97	0.995	0.999	0.997	0.911	575	0	93	16
1000	yes	1, group:s+v	0.969	0.958	0.996	0.999	0.998	0.939	800	0	132	15
1000	yes	25, group:s+v	0.968	0.954	0.995	0.999	0.996	0.935	799	0	132	16
1000	no	1, group:s+v	0.966	0.955	0.995	0.999	0.995	0.934	800	1	131	15
1000	no	1, com:cor	0.966	0.958	0.995	0.999	0.995	0.927	798	1	131	17
1000	yes	1, com:cor	0.971	0.961	0.996	0.999	0.998	0.92	795	0	132	20
1000	yes	25, com:cor	0.971	0.961	0.996	0.999	0.998	0.92	795	0	132	20
1000	yes	1, slices	0.956	0.939	0.993	0.999	0.994	0.91	11096	6	1842	314
1000	no	1, slices	0.953	0.94	0.993	0.999	0.995	0.905	11087	13	1835	323
1000	yes	1, com:sag	0.956	0.943	0.991	0.999	0.995	0.893	790	2	130	25
1000	yes	25, com:sag	0.956	0.943	0.991	0.999	0.995	0.893	790	2	130	25
1000	no	1, com:sag	0.948	0.934	0.991	0.998	0.994	0.883	787	2	130	28
2600	yes	25, group:s+v	0.956	0.899	0.956	0.996	0.986	0.816	1192	1	117	50
2600	yes	1, group:s+v	0.953	0.894	0.954	0.995	0.986	0.813	1191	1	117	51
2600	no	1, group:s+v	0.953	0.9	0.958	0.996	0.986	0.81	1190	1	117	52
2600	yes	1, com:sag	0.928	0.868	0.948	0.994	0.983	0.797	1185	1	117	57
2600	no	1, com:sag	0.922	0.859	0.931	0.993	0.978	0.797	1185	1	117	57
2600	yes	25, com:sag	0.928	0.868	0.948	0.994	0.983	0.797	1185	1	117	57
2600	no	1, slices	0.921	0.85	0.94	0.994	0.98	0.78	16504	16	1636	884
2600	yes	1, slices	0.919	0.849	0.933	0.993	0.98	0.78	16505	18	1634	883
2600	no	1, com:cor	0.915	0.846	0.951	0.994	0.97	0.764	1174	2	116	68
2600	yes	1, com:cor	0.925	0.859	0.946	0.993	0.981	0.754	1172	3	115	70
2600	yes	25, com:cor	0.925	0.859	0.946	0.993	0.981	0.754	1172	3	115	70
all	yes	25, group:s+v	0.958	0.931	0.987	0.998	0.992	0.885	2747	1	367	86
all	yes	1, group:s+v	0.957	0.932	0.987	0.998	0.993	0.884	2746	1	367	87
all	no	1, group:s+v	0.958	0.934	0.987	0.998	0.992	0.879	2743	2	366	90
all	yes	1, com:sag	0.946	0.919	0.981	0.998	0.99	0.862	2729	3	365	104
all	yes	25, com:sag	0.946	0.919	0.981	0.998	0.99	0.862	2729	3	365	104
all	no	1, com:sag	0.941	0.91	0.979	0.997	0.987	0.856	2724	3	365	109
all	no	1, slices	0.939	0.908	0.981	0.997	0.988	0.856	38118	34	5118	1544
all	yes	1, slices	0.939	0.905	0.98	0.997	0.988	0.855	38092	27	5125	1570
all	no	1, com:cor	0.938	0.909	0.983	0.998	0.988	0.852	2722	4	364	111
all	yes	1, com:cor	0.942	0.912	0.983	0.997	0.988	0.849	2717	3	365	116
all	yes	25, com:cor	0.942	0.912	0.983	0.997	0.988	0.849	2717	3	365	116

Table 6.19.: The effect of different groupings of CNN classification results to form the final classification on performance measures for a model of the *scratch_22* architecture. The *aug* column indicates whether data augmentation was used for testing. The sample column splits the performance measures into different classification sampling strategies. The number in the sample column indicates how many augmented (or non-augmented) versions of each image were used and the abbreviations after the comma show the grouping on the volume level: *com*: centre of mass, *sag*: sagittal, *cor*: coronal, *group:s+v*: classification label averaged grouped by subject and volume (possibly across augmentations). Results are rounded to three digits and sorted by b-value using the b-value specific model-independent performance rank see text).

b	sp95r	H	MCC	AP	AUROC
0	0.96 [0.95, 0.96]	0.90 [0.89, 0.90]	0.80 [0.79, 0.80]	0.959 [0.957, 0.962]	0.993 [0.993, 0.994]
400	0.9978 [0.9976, 0.9979]	0.992 [0.991, 0.993]	0.93 [0.92, 0.93]	0.9960 [0.9957, 0.9962]	0.9994 [0.9994, 0.9995]
1000	1.00 [0.99, 1.00]	0.960 [0.958, 0.961]	0.91 [0.90, 0.91]	0.9939 [0.9936, 0.9943]	0.9989 [0.9988, 0.9990]
2600	0.978 [0.977, 0.979]	0.92 [0.92, 0.93]	0.76 [0.75, 0.78]	0.931 [0.929, 0.933]	0.992 [0.992, 0.993]
all	0.988 [0.987, 0.988]	0.940 [0.939, 0.942]	0.85 [0.84, 0.85]	0.980 [0.979, 0.981]	0.9970 [0.9968, 0.9971]

Table 6.20.: Comparison of mean and 95% confidence intervals of 5 performance metrics. Performance was evaluated for all sampling schemes listed in table 6.19 across 14 models of the architectures *vgg16_22333*, *scratch_2*, *scratch_22*, *scratch_223*, *scratch_2233*, *scratch_2233d*, and *scratch_2233d_noaug*.

6.4.1.1. Metric selection

For model performance comparisons, it is helpful to find one or two representative metrics and testing schemes instead of discussing results such as table 6.19 across model instances. For this purpose, a representative sample of 7 model architectures, trained twice on 34 subjects yielding 14 different models was selected and analysed across test sampling schemes with the goal of finding a combination of metric and test method that is representative.

The selected architectures span the range from shallow neural networks (*scratch_2*) to deeper networks (*scratch_2233d*) to the model (*vgg16_22333*) and deep networks trained without data augmentation (*scratch_2233d_noaug*).

Comparing metric values Table 6.20 lists the mean performance values and 95% confidence intervals of all 14 models evaluated across the 11 test sampling schemes listed in table 6.19 for 5 performance measures: specificity at 95% recall (sp95r), Matthews correlation coefficient (MCC), H measure, average precision (AP) and area under the ROC (AUROC). Confidence intervals were calculated using the t-statistic and logit transform, treating all models as independent.

The average performance values in table 6.20 differ across b-values presumably due to varying classification performance but also due to b-value dependent degrees of class imbalance and how each metric weights misclassification costs under these circumstances (compare fig. 5.3). Also the relative spread of performance values across models and testing scenarios for a given b-value is metric dependent. In general, AUROC and AP require up to four digits of precision for reporting the 95% confidence intervals while MCC CI values vary in the second digit. Furthermore, the performance ranking across b-values is also metric-dependent. For instance, the classification of the b=0 slices has the lowest H-value but MCC, AP and AUROC rank the 2600 shell performance significantly worse than the b=0 performance.

Consensus between metrics If it is possible to assign specific costs to misclassification of artefacts or non-artefacted data, one should chose a performance metric that takes this into account [Hand, 2006]. However, it is not trivial to determine the effect motion artefacted volumes have on any diffusion processing pipeline and its outcome. This

cost likely depends on the number and quality of image data close by in q-space and the robustness of the processing pipeline. Therefore, an approach could be to chose the metric that ranks different models in a similar order as most other metrics. This assumes the consensus between metrics being a proxy for true performance but also introduces a stability in the results obtained - independent of the application's true misclassification cost.

I am not only interested in the agreement of what the best model is across metrics but also in finding a metric that is representative for the task of ranking models. However, selecting a metric based on the agreement with the consensus rank might be at the expense of selecting a metric with lower sensitivity for model-selection evaluation. If most metrics considered rank consistently, but differently, a single metric that is less sensitive but happens to align better with the average ranking would, it appear as superior. If a range of models perform at different levels, an ideal metric would rank them in that order.

For instance, assume a set of models with performance ranks $\mathbf{r} = [1, 3, 4, 2]$, with 1 being the best performing model. If a given metric M yields the performance values $[0.7, 0.8, 0.9, 0.8]$ it could assign the 'average' ranking $\mathbf{r}^M = [1, 2.5, 4, 2.5]$ or 'minimum' ranking $\mathbf{r}^M = [1, 2, 4, 2]$. The mean average distance (MAD) from the true ranking is independent of the ranking scheme (1 in both cases). For comparison, a reversed order between two consecutive models would yield a distance of 2. However, the consensus rank is affected by the choice of ranking method and whether the mean or median across models is used. I chose the median since stability is the criterion I am trying to assess, and average ranking, as it keeps the sum across ranks constant.

Rankings of the N=14 models were calculated for each combination of b-value, test sampling method, and test metric (H, AP, AUROC, sp95r and MCC). The consensus ranking is defined as the median ranking across metrics. A representative metric has little deviation d from the consensus irrespective of the test sampling scheme:

$$d = \left\langle \frac{1}{N} \sum_{i=1}^{N} \mathrm{abs}(r_i - \mathbf{r}_i^M) \mid \text{b, test sample method} \right\rangle_{\text{b, test sample method}}$$

The highest agreement in model rank with the consensus when averaged over all b-values, test augmentation and sample methods has AUROC (1.51), followed by AP (1.60), H (1.74) and sp95r (1.77). MCC disagrees with the median rank far more often, on average by 3.03 positions. sp95r produces tied performance values in 26% of rankings, followed by AUROC with 2.4%, MCC with 1.8%, and H and AP with each 1.0% tied values on average. Hence, sp95r is far less sensitive for ranking models.

Figure 6.9 shows the deviation of the model ranking for the 14 models split up by b-values (columns) and metrics (rows). A boxplot with small width indicates strong agreement of the metric with the consensus in the respective b-value and sampling scenario. Small whiskers in presence of points with large (positive or negative) rank difference values indicate metric instability. In the case of MCC, the 25th and 75th percentile (box edges) are large, showing that this metric ranks systematically differently than the consensus. Rankings performed on the slice-level and using only the centre of mass

slices show higher disagreement between metrics. Hence model selection based on those measures is likely less reliable.

Figure 6.9.: Difference between the consensus model ranking and the model ranking using a single metric for ranking (columns). Model rankings are calculated with fixed b-value and test sample for each metric independently. The consensus rank for each model is the median rank across metrics.

6.4.1.2. Effect of test data sampling methods

So far, all comparisons were made with the 11 test sampling method combinations (test image augmentation, groupings of slices, 25 repetitions using augmentations) kept fixed and averaged over. However, those test data sampling methods give insights into the classifier and metric behaviours for different test domains. Analysing the model-independent

performances under these different scenarios allows gaining insights into CNN classifier performance and stability under different classification conditions:

Testing using all slices independently is a measure for average slice-wise performance. Using only the central slice of the coronal projections (sample = *com:cor*) or that of the sagittal projections (sample = *com:sag*) is a per-volume measure derived from single slices. In contrast, averaging the classification outcome of all 14 slices per subject and volume (sample = *group:s+v*) combines multiple points of view of the anatomy into a final classification. The effect of augmenting the test images can be observed by comparing classifier performance without augmentation, with augmentation on a single augmented sample or with augmentation and 25 augmentation repeats for each slice.

Figure 6.10 shows the performance ranking of test sampling methods across models. The sampling methods are ranked for each model and metric independently. Shown is the median (top row) across metrics for all models. For b=0, the sagittal centre of mass slice gives overall highest classification performance and is superior to the coronal slice and the average across slices. For all other b-values, the test sampling using each slice with classification results averaged across all slices yields the highest performance, especially in the high b-value regime. Using 25 augmentations of all slices further improves performance. These findings are also valid for the two models trained without data augmentation (*scratch_2233d_noaug*, data not shown). The bottom row shows the disagreement (MAD) between metrics on the ranking of each sampling method. The metrics agree most on the ranking of the overall best performing method using test data augmentation and 25 repetitions of each slice. In general, metrics agree more in the scenario of averaging classifications across multiple slices (*groups:s+v*).

In contrast to threshold-based measures, the integral metrics H-measure, AUROC and AP are designed to be more stable in the presence of changing test conditions. This is apparent in the number of unique rankings these metrics assign to the same model tested under 11 slightly different conditions. AUROC yields on average 4.9 distinct ranks, H-measure 5.2 and average precision 5.6. This is in contrast to MCC and sp95r, which yield 6.6 and 7.2 ranks, respectively.

6.4.1.3. Conclusion

Specificity at 95% recall is sensitive to changes in test conditions but if they are controlled for, agrees well with the consensus of the other metrics. AUROC, AP and H rank models more consistently and are the least impacted by changes in test conditions. Rankings based on MCC are the most distinct from rankings generated using the other metrics and are slightly more sensitive to changes in test conditions than the integrative metrics. H-measure agrees well with the consensus but uses an explicitly defined weighting of misclassification cost that depends on the class imbalance. Model rankings produced via the AUROC measure are the most representative and AUROC rankings are the least sensitive to changes in test conditions. MCC is the most dissimilar from all other metrics. Hence further discussions report H-measure, AUROC and MCC but use AUROC for ranking models.

The non-augmented sagittal slice located in the centre of mass seems to be ideal for

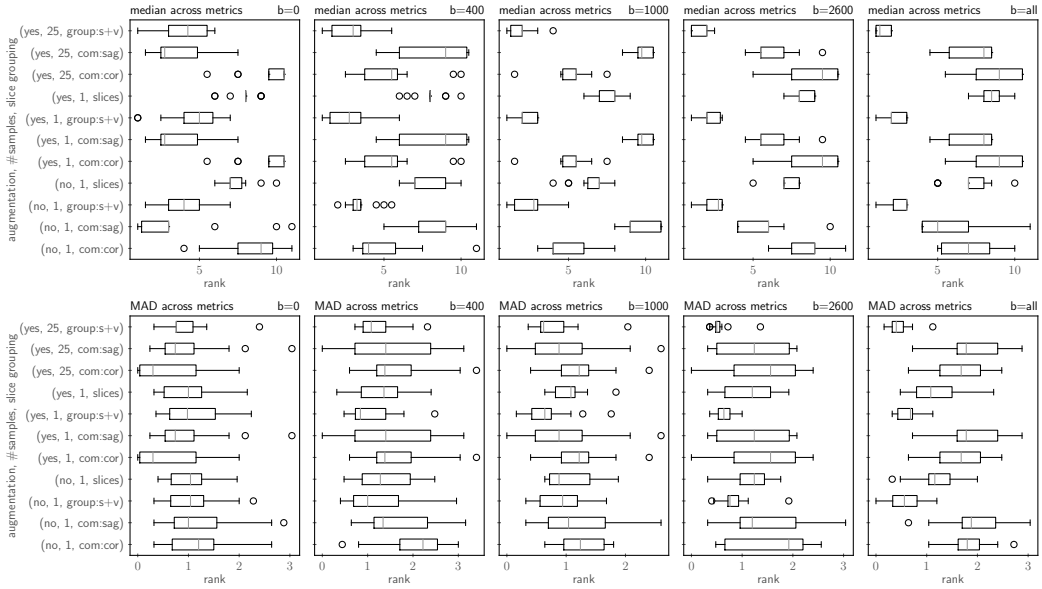


Figure 6.10.: Performance rankings of different test data sampling methods across 14 models. Top: Median rank across the metrics H, AP, AUROC, sp95r and MCC. Bottom: Disagreement between metrics on the sampling rank (mean average deviation).

classification in the absence of diffusion weighting and augmentation and the other slices degrade classification performance. Classifier performance for the other shells is best when classification labels are averaged across all coronal and sagittal slices. Furthermore performance can be improved by augmenting 25 versions of the test images. This effectively simulates a pooling across slightly different classifiers because artefacts are presented at different locations of the neural network and therefore processed differently. For all further analysis I choose pooling classification labels of 25 augmentations across all slices as it yields rankings that are similar for most metrics, indicating high generalisability of rankings.

6.4.2. Data properties and training parameters

6.4.2.1. Training data size and augmentation

Training data size For models of the *scratch_2233d* architecture and models reusing parts of the pre-trained VGG16 model, model performance increases consistently with the number of training subjects (see table 6.21). Also the spread in performance between different training runs decreases with increasing number of subjects.

While it is plausible that networks trained on fewer subjects are less consistent between training runs, the difference of the spread in table 6.21 is likely biased by the lower sample variability in the high subject regime. For models trained on fewer subjects, the training subjects were selected randomly and therefore training runs show higher variability. The

number of subjects	AUROC	AUROC ± 95% CI	H ± 95% CI	MCC ± 95% CI
scratch_2233d				
2	0.9976			
15	0.9978			
30	0.9979			
34	0.9979			

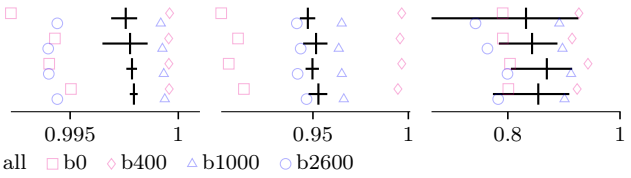


Table 6.21.: Effect of the size of the training data on classification performance of models of the *scratch_2233d* architecture. Models were trained on 2, 15, 30 and 34 subjects. All numeric values correspond to the joint testing of all b-values (b=*all*) for which the average and confidence intervals across 3 training runs are shown. The performance for individual b-values is shown as their average across training runs omitting error bars for clarity.

anatomical variability between training runs is drastically reduced in the case of 34 subjects as only 2 subjects' role can be swapped between training and validation data. The order of samples presented to the network was shuffled randomly before each epoch and therefore each run, which leads to different model updates after each training batch and hence different models even when trained on identical data. Sample selection bias can only partially be accounted for by shuffling the training data before each run and a fair comparison would require more training data or bootstrap resampling, which requires training hundreds of models.

Augmentation All models were trained on multiple slices from each volume, which can be viewed as a form of natural data augmentation that preserves the image characteristics but varies image features. Slices from the same volume are not independent but they likely provide additional information that shares the volume's artefact characteristics.

Training image distortion further increases the effective size of the training data. The goal is to augment training in ways that preserves the characteristics necessary for classification but offers additional information to the network. For instance, shifting the image content can be used to learn models that are less dependent on the input image location by learning feature representations at each input location - if the network parameter capacity allows it. However, if the variability in the training data is sufficient then there is no need to augment the data. In that case artefacts caused by image augmentation can degrade classification performance. This trade-off depends on the amount and quality of training data, the problem characteristics, and the network's properties. Image rotation, for instance is a plausible augmentation for object detection but unlikely to help detecting stripe artefacts that are always aligned with the pixel grid. For the latter application, even regridding the images, which blurs fine textures and smooths edges could degrade performance.

Table 6.22 lists the experiments performed to investigate the effect of training augmentation on classifier performance for models of the *scratch_2233d* architecture, trained on 2 and 34 subjects. The training augmentation components investigated are: no augmen-

tation, horizontal shift (*hshift*), horizontal and vertical shift (*shift*), horizontal flipping (*flip*) and zooming. For computation resource reasons, not all combinations were tested.

The number of subjects impacts performance to a much higher extent than the augmentation scheme. Across metrics, models trained without augmentation perform no worse than models trained with full augmentation. Models trained on 34 and 2 training subjects using *hshift+flip+zoom* image augmentation (the default) perform comparatively consistently across training runs. The relative insensitivity to image augmentation can be interpreted as a sign of an easy classification problem and the presence of sufficient variability in the data [Perez, Wang, 2017].

In the case of 34 subjects, horizontal shifting without zooming or flipping, has the highest AUROC value. This is the augmentation scheme that does not blur in axial direction. In the case of 2 training subjects, *hshift+flip+zoom* yields the highest AUROC value. Variability between training runs is too high to make definite statements, but this suggests that zooming in combination with flipping is beneficial for small datasets, but degrades performance in the full dataset likely due to blurring. Even the full training set can profit from augmentation if it conserves texture characteristics perpendicular to the (axial) plane direction.

6.4.2.2. The effect of class imbalance and remedies

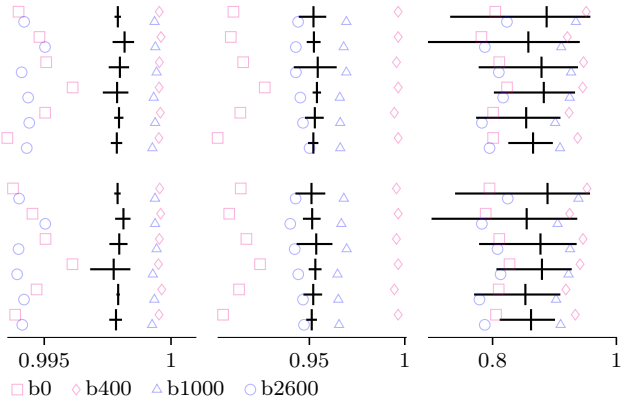
Medical data is rarely balanced in the prevalence of the labels and indeed our dataset contains only about 15% artefact positive labels. Furthermore, the number of volumes in each b-value is very different, which results in an overall small number of artefacted b=0 samples compared to artefacts in the other shells. If a classifier learns a b-value independent criterion for artefacts then this within-class imbalance should not matter. As discussed earlier, two easy to implement strategies that aim at improving classifier performance in the presence of class imbalance are minority class oversampling and cost function weighting based on class-prevalence.

During a training epoch, a neural network is presented with each training sample at least once. In the case of oversampling the minority class[5], an epoch contains multiple versions of the minority class but each sample of the majority class once. It is plausible that a network learns faster and better if it receives a balanced mix of positive and negative examples. However, the minority samples are very similar repetitions of the same data (or identical for non-augmented training) and might cause the network to overfit to their specific characteristics and anatomy. Hence there might be a trade-off in oversampling that depends on the data, the network and the classification problem.

Minority class oversampling could be implemented on the epoch-level, balancing the overall number of samples from each class, or on the batch-level, which is used here, mainly for programming convenience reasons but might have an impact on the *custom* architecture due to its batch normalisation layers. I investigate the effect of no over-

[5]Using data-augmentation and training until convergence, the only differences between oversampling the minority class and undersampling the majority class are the number of steps in each epoch and how often each sample of the majority class is seen due to random selection in the undersampling case.

augmentation, 34 subjects	AUROC	AUROC ± 95% CI	H ± 95% CI	MCC ± 95% CI
test augmentation				
none	0.9979			
hshift	0.9982			
shift	0.9980			
shift+flip	0.9979			
hshift+flip+zoom	0.9979			
shift+flip+zoom	0.9979			
no test augmentation				
none	0.9979			
hshift	0.9981			
shift	0.9980			
shift+flip	0.9977			
hshift+flip+zoom	0.9979			
shift+flip+zoom	0.9978			



augmentation, 2 subjects	AUROC	AUROC ± 95% CI	H ± 95% CI	MCC ± 95% CI
test augmentation				
none	0.9975			
hshift	0.9972			
shift	0.9974			
shift+flip	0.9973			
hshift+flip+zoom	0.9976			
shift+flip+zoom	0.9974			
no test augmentation				
none	0.9972			
hshift	0.9969			
shift	0.9973			
shift+flip	0.9972			
hshift+flip+zoom	0.9975			
shift+flip+zoom	0.9974			

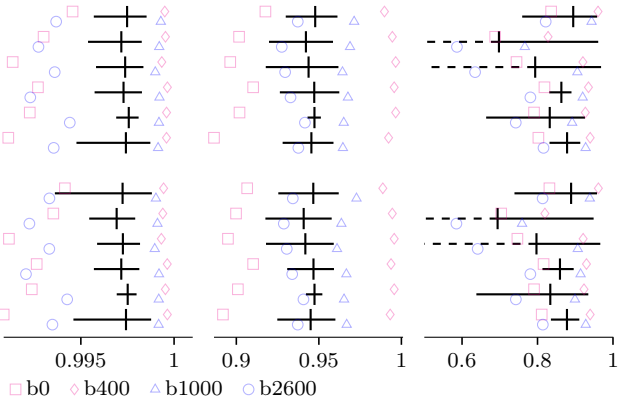


Table 6.22.: Effect of training augmentation methods on classification performance of models of the *scratch_2233d* architecture trained on 34 subjects (top) and trained on 2 subjects (bottom) evaluated without test image augmentation and with *shift+flip* test augmentation. *hshift* stands for horizontal shift, *shift* for both vertical and horizontal shift, *flip* for horizontal flipping (anatomical left-right or anterior-posterior), and *zoom* for scaling the image. Results obtained using test augmentation are very similar to those without test image augmentation. The number of subjects impacts performance at a much higher degree than the augmentation scheme.

sampling (*no bal*), oversampling images of the minority label (*l bal*), and additionally balancing b-values in each batch (*l+b bal*), which was the default for all other training experiments.

Besides oversampling, up-weighting the cost (and parameter update) of the minority class relative to the majority class, "incentivises" the network to focus on samples of the minority class. This can be seen as a class-dependent regularisation of the step size, increasing the step size for the minority and lowering it for the majority class. Depending on the cost-landscape, a network could end up in the same or a very different cost-function minumum.

Similar to sample weighting in logistic regression [King et al., 2001], I implemented sample weighting based on the prevalence of labels ($\mathbf{l}$) in the training set (*l weight*):

$$w(l) = \frac{N_{\text{samples}}}{2 \sum_{i=1}^{N_{\text{samples}}} I(\mathbf{l}_i = l)} \tag{6.3}$$

and based on the distribution of labels and b-values ($\mathbf{b}$) in the training set (*l+b weight*):

$$w(l, b) = \frac{N_{\text{samples}}}{8 \sum_{i=1}^{N_{\text{samples}}} I(\mathbf{l}_i = l) I(\mathbf{b}_i = b)} \tag{6.4}$$

Note that the vectors $\mathbf{l}$ and $\mathbf{b}$ contain all ground truth labels and b-values of a training epoch. Hence, the weight is normalised on the epoch level. Random sampling from the training data causes stochastic fluctuations of weight each batch carries. To regularise potential overshooting due to high prevalence of rare samples in a batch, I implemented a third weighting scheme that uses eq. (6.4) but normalises the weight in each batch to sum to the batch size (*l+b weight BN*).

To sum up, all tested combinations of sampling and weighting are: no oversampling or weighting (*no bal*), using oversampling of the minority label (*l bal*), additionally balancing b-value within-class imbalance (*l+b bal*), not oversampling but weighting by label distribution (*l weight*), weighting by label and b-value distribution without regularisation (*l+b weight*) and normalised on the batch-level (*l+b weight BN*). These methods were tested on models of the architectures *scratch_223*, *scratch_2233d*, and *custom*, each trained 3 times.

The performance of models trained on non-balanced batches varied much more. Not balancing the data yielded a high number of failed models: none of the 3 training attempts was successful for the combinations *scratch_223* with *no bal, l weight* and for *custom* with *no bal, l weight*, *no bal, l+b weight*, or *no bal, l+b weight BN*. The *custom* architecture was most unstable on unbalanced batches and produced only 2 useful models out of the 12 training attempts. This is likely related to the batch normalisation layers.

For the *scratch_223* architecture, balancing labels and b-values is optimal, however the other architectures performed equally well or better when trained on batches with balanced labels or in the absence of balancing. However, only 1 out of 3 non-balanced models were successfully trained for the custom and *scratch_2233d* architectures.

Not enough models using label weighting trained successfully to make detailed statements about the different weighting methods. Of the 9 attempts, only 2 models trained

sampling / weighting (n)	AUROC	AUROC ± 95% CI	H ± 95% CI	MCC ± 95% CI
scratch_223				
l+b bal (3)	0.9981			
l bal (3)	0.9979			
no bal (2)	0.9972			
no bal, l+b weight (1)	0.9952			
no bal, l+b weight BN (2)	0.9978			
scratch_2233d				
l+b bal (3)	0.9979			
l bal (3)	0.9980			
no bal (1)	0.9980			
no bal, l+b weight (2)	0.9945			
no bal, l+b weight BN (2)	0.9979			
no bal, l weight (1)	0.9976			
custom				
l+b bal (3)	0.9982			
l bal (3)	0.9982			
no bal (1)	0.9982			
no bal, l weight (1)	0.9981			

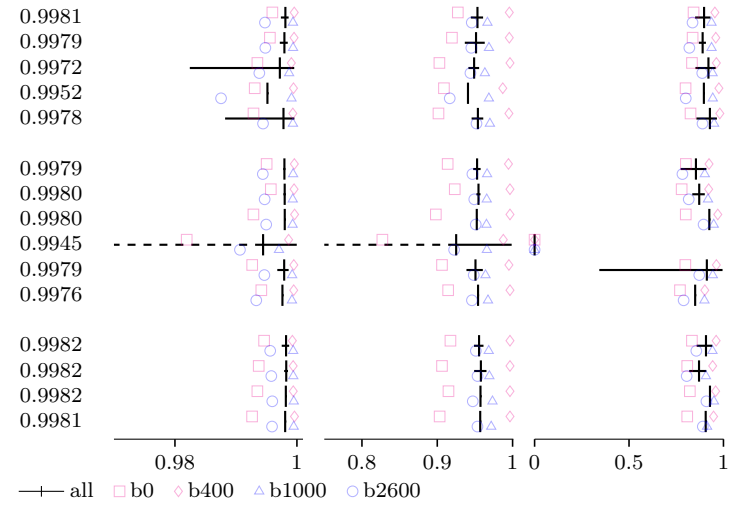


Table 6.23.: Effect of class imbalance remedies on classifier performance for models of 3 different architecture. All models were trained on 3 time using different training data but only the n models were used for the analysis as the others did not achieve an AUROC greater than 0.8 on the validation set. The absence of a combination means that none of the 3 models was trained successfully. Error bars exceeding the plotted range are shown with dashed lines.

successfully using label weighting. Label and b-value weighted training was more stable. However, in all architectures, weighting decreases model performance compared to non-weighted unbalanced training. Batch normalised weighting improves performance over unnormalised weighting.

6.4.3. Network architectures and transfer learning

6.4.3.1. Depth, number of free parameters, filter dimensionality

The networks investigated are only a small subset of the networks used for computer visions and likely not representative to make general statements about the required number of layers and network capacity. However, within the constraints of VGG-style networks, a few general trends emerge.

Table 6.24 shows the performance of all models trained on the full dataset and using a common training strategy sorted by the total number of trainable parameters and sorted by the number of layers. Overall, the performance across architectures is very similar but high-parametric models tend to be less stable. For the networks trained from scratch, a total number of 7 or 10 convolution layers (*scratch_223*, *scratch_2233*) is beneficial. The *scratch_2233* model performance can be improved slightly by quadrupling the number of filters in the first layer, at the expense of training stability and drastic increase in computation time.

The *custom* to *custom5* architectures perform very similarly. Convolution filter shape and size seem to matter very little. Again, the highest parametric model has the highest training-run variability. All *custom* models achieve scores equal to or higher than any VGG-like models trained from scratch.

Compared to the training method, the choice of architecture plays a minor role on performance. The overall two best model architectures *custom*, with 54 thousand free parameters, and *vgg16_223nopool*, which requires training 394 thousand parameters, achieve an AUROC of 0.9982.

6.4.3.2. Transfer learning from pre-trained VGG16 network

The power of transfer learning is to reuse a pre-trained network to extract features learned in a different domain. Hence, in the new domain, it provides a set of features without the need for supervised learning. All the training data can be used to combine these features to form a prediction, which makes transfer learning appealing for data-scarce domains.

Networks that reuse parts of the VGG16 network weights and has less than 8.5 thousand parameters to recombine those features performed worse than the networks with higher parameter count (see table 6.24). The worst-performing model is the network that recycles the first convolution block of the VGG16 network (*vgg16_2*) and happens to be the network with the least trainable parameters. However, this network also uses global average pooling after only two convolution layers (see table 6.10). Hence, its receptive field is at most 5x5 and the final 3 classification layers do no have access to any information about the location of the extracted feature vectors.

model	AUROC	AUROC ± 95% CI	H ± 95% CI	MCC ± 95% CI
sorted by # parameters				
custom (54k)	0.9982			
custom3 (58k)	0.9981			
custom2 (77k)	0.9981			
custom4 (78k)	0.9981			
custom5 (166k)	0.9981			
scratch_223 (52k)	0.9981			
scratch_22 (53k)	0.9979			
scratch_2 (101k)	0.9979			
scratch_2233 (132k)	0.9981			
scratch_2233d (132k)	0.9979			
scratch_22333d (495k)	0.9979			
scratch_2233d32 (2m)	0.9980			
vgg16_2 (1.3k)	0.9976			
vgg16_22 (2.4k)	0.9978			
vgg16_223 (4.4k)	0.9980			
vgg16_2233 (8.5k)	0.9981			
vgg16_22333 (49k)	0.9981			
vgg16_2233nopool (197k)	0.9981			
vgg16_223nopool (394k)	0.9982			
vgg16_22nopool (788k)	0.9980			
vgg16_2233f3 (7.1m)	0.9981			
vgg16_f22333 (14.8m)	0.9982			
sorted by # layers				
custom (6)	0.9982			
custom2 (6)	0.9981			
custom3 (6)	0.9981			
custom4 (6)	0.9981			
custom5 (6)	0.9981			
scratch_2 (5)	0.9979			
scratch_22 (7)	0.9979			
scratch_223 (10)	0.9981			
scratch_2233 (13)	0.9981			
scratch_2233d (13)	0.9979			
scratch_2233d32 (13)	0.9980			
scratch_22333d (16)	0.9979			
vgg16_2 (5)	0.9976			
vgg16_22 (7)	0.9978			
vgg16_22nopool (7)	0.9980			
vgg16_223 (10)	0.9980			
vgg16_223nopool (10)	0.9982			
vgg16_2233 (13)	0.9981			
vgg16_2233nopool (13)	0.9981			
vgg16_22333 (16)	0.9981			
vgg16_2233f3 (16)	0.9981			
vgg16_f22333 (16)	0.9982			

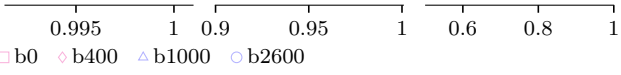


Table 6.24.: Performance of model architectures trained from scratch and using transfer learning. All models were trained on 34 subjects with data augmentation. Architectures are sorted by the number of trainable parameters (top half) and by the number of fully connected layers (bottom half). 1 of the 3 models of the *scratch_2233* and *scratch_2233d32* architectures failed to train and are excluded from the plots.

The best performing VGG16-derived network, *vgg16_223nopool*, reuses the first 3 blocks and retains the spatial information for the final classification. This model has 394 thousand trainable parameters but trains remarkable robustly without using regularisation compared to models trained from scratch using much fewer parameters. The *vgg16_2233* model performs as well or better than any *scratch** model, despite it having only 8.5 thousand parameters. Overall, there is no benefit in fine-tuning the last convolution block or reusing deeper convolution layers, unless all layers are fine-tuned (*vgg16_f22333*). This is most beneficial for the b=0 shell but makes training very time-consuming and model performance variable.

Hence, if computational resources are limited, it is best to reuse only 3 to 4 blocks. This makes training and inference much faster and less memory demanding.

Table 6.25 compares the performance of models trained on 34 and on 2 subjects. Using only 2 subjects, the full VGG16 network with re-trained last dense layers achieves the best AUROC of 0.9978 and is the most stable of the models trained on 2 subjects. The *vgg16_22* network performs surprisingly well, given that is uses only 4 convolution layers followed by spatial pooling. However, the network's capability is limited as it barely improves when trained on 34 instead of 2 subjects. Fine-tuning and training classifiers on a higher spatially resolved feature map degrades performance presumably because the training data can not support the high number of parameters.

6.4.3.3. Architecture versus augmentation ensembles

Pooling multiple decisions using test data augmentation improves model performance. However, model architectures or instances could also be pooled to boost performance [Freund, Schapire, 1997]. This technique is usually applied to "weak" learners that are easy to train but, provided the compute resources, can also be applied to neural networks. Provided that all classifications are better than chance, performance is expected to be higher, the more dissimilar the classifiers are [Hastie, Friedman, Tibshirani, 2009a].

Table 6.26 shows a comparison of performance achieved by averaging predictions from 25 randomly chosen models of certain model architectures, evaluated on a single augmentation ('model ensemble'), and evaluated for each model independently but using 25 augmented test labels.

Models of each category were chosen randomly across training runs from all models trained on 34 subjects and on all b-values using *l+b bal* sampling. For both ensemble methods, the selected models were the same. For the model ensemble, confidence intervals were calculated by sampling 25 different augmentations and in the augmentation sampling, they represent the spread across models. Confidence intervals do not account for the model selection variability. The *custom** architecture ensemble has only 15, not 25, different models. Therefore, the *custom** test augmentation ensemble was also calculated using only 15 augmentations.

Except for the ensemble that consists only of *scratch** models, performance evaluated on all b-values for model ensembles is on par or higher than the average performance for individual models using test augmentation. However, the variability in performance of the model ensembles is higher for individual b-values.

model / subjects	AUROC	AUROC ± 95% CI	H ± 95% CI	MCC ± 95% CI
vgg16_2 (1.3k)				
34	0.9976			
2	0.9972			
vgg16_22 (2.4k)				
34	0.9978			
2	0.9977			
vgg16_223 (4.4k)				
34	0.9980			
2	0.9974			
vgg16_2233 (8.5k)				
34	0.9981			
2	0.9974			
vgg16_22333 (49k)				
32	0.9981			
2	0.9978			
vgg16_2233nopool (197k)				
34	0.9981			
2	0.9970			
vgg16_223nopool (394k)				
34	0.9982			
2	0.9969			
vgg16_22nopool (788k)				
34	0.9980			
vgg16_2233f3 (7.1m)				
34	0.9981			
2	0.9975			
vgg16_f22333 (14.8m)				
34	0.9982			
2	0.9976			

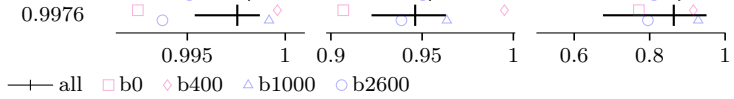


Table 6.25.: Performance of transfer learning model architectures with last layers trained from scratch using 34 and 2 subjects. The number of free parameters (using random initialisation or initialised with the VGG16 weights in the case of fine-tuned models) are shown in brackets.

ensemble / architecture	AUROC	AUROC ± 95% CI	H ± 95% CI	MCC ± 95% CI
model ensemble				
custom*	0.9982			
scratch*	0.9976			
scratch* custom*	0.9981			
scratch* custom* vgg16*	0.9980			
scratch_223*	0.9980			
vgg16*	0.9981			
test augmentation ensemble				
custom*	0.9981			
scratch*	0.9979			
scratch* custom*	0.9980			
scratch* custom* vgg16*	0.9979			
scratch_223*	0.9980			
vgg16*	0.9980			

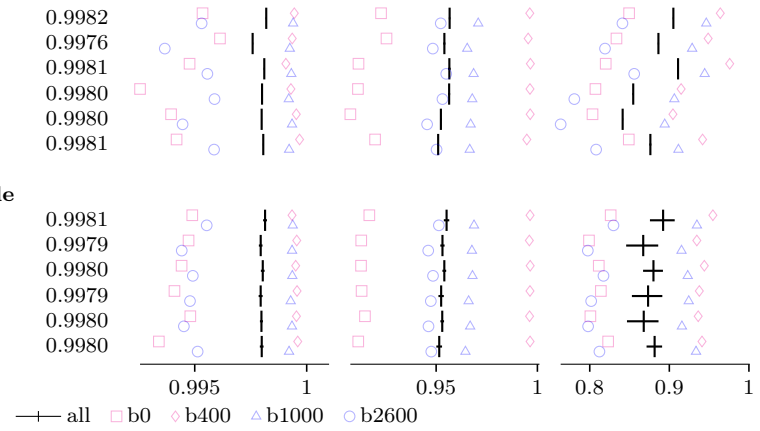


Table 6.26.: Comparison of performance of model ensembles consisting of 25 models evaluated on one random augmentation each (top) with individual models using ensembles of test augmentations (bottom).

6.4.4. b-value specific training: domain adaptation and within-class structure

To investigate the within-class structure of the data, models of the *scratch_2233d* architecture were trained on subsets of the b-value range. Table 6.27 lists the network performance split by b-value for all models trained on 34 (top) and 2 (bottom) training subjects.

The *scratch_2233d* model trained on the b=0, 400, and 1000 shells and all subjects achieves the highest AUROC and H test score on the shells it was trained on, surpassing the model trained on all data. This model performs surprisingly well on the b=2600 shell, where it has a higher AUROC than the model trained on all data but lower H score and MCC. The best model on the b=2600 shell is the model that used only that shell to train.

That a model that has not been trained on the highest shell performs relatively well suggests that the networks have learned a feature representation that is robust across b-value and therefore image contrast changes. Usually, training on more data is beneficial for performance. However, including all shells in the training data decreases test performance in the highest shell compared to the model trained only on the b=2600 shell. Using all data is not beneficial or decreases test performance on the lower shells as well. This suggests that the networks learn competing criteria for the lower and the highest shell.

As diffusion weightings are interspersed within a scan (see section 6.3.1), if the movement of a baby is independent on the b-value of the volume, and if motion affects the image quality of all b-values equally, then the ratio of good to bad volumes should be equal for all b-values. Under these assumptions, the training and test data are biased:

91.3% of the b=2600 volumes are labelled as usable volumes, but only 88.1%, 86.4% and 86.1% of the b=0, 400 and 1000 shells are 'acceptable' by my standards. For the second observer, the ratios are similarly skewed, albeit with overall slightly more lenient criteria.

The *scratch_2233d* network labelled 82.4%, 85.9%, 83.9% and 82.9% of the test set volumes as usable. Overall, the network tends to reject more volumes in any b-value, especially in the b=0 and b=2600 shell. The network trained only on the b=0 shell has the same low acceptance ratio for that shell. Lacking definite ground truth labels, it is not possible to prove that the networks have learned a less biased criterion but it seems plausible: human observers know that b=0 samples are essential for distortion correction and motion artefacts are likely to "average out" on that shell due to the lack of diffusion weighting. The reduced acceptance rate in the b=0 classifications compared to the average classification acceptance rate might be caused by spin history artefacts, which can manifest as stripe-patterns in the sagittal and coronal projections and are most visible in the b=0 shell. This is in line with the observation that the *scratch_22333d* network has learned to focus its "attention" on the edge of the brain in the b=0 shell (see section 6.7.2.2) but uses the full brain and parts of the reconstruction mask in the lower shells.

The network trained on the b=2600 shell has an acceptance rate of 87.5% in that shell, indicating that it is possible for the networks to better adjust to the human rating criterion but this rate goes down to 86.8% for the network trained on all data and even further down to 82.9% for the network trained on the 3 lower shells, which have on average 86.4% good volumes. Also, the performance on the b=2600 shell seems to be independent of the number of parameters (see table 6.24). Hence it is not the lack of network capacity that is limiting learning a good representation. The performance of *scratch_2233* on the b=0 and b=2600 shells is lower than for the models trained with dropout regularisation. This is reversed (using the AUROC) for the other shells. Lacking definite ground truth, it is left to the reader to speculate, whether dropout regularisation leads to learning more representative features or to decreased performance due to over-reliance on more distributed features.

When trained on only 2 subjects, performing b-value specific training or a combination of the lower 3 shells produces 9 best results across AUROC, H-measure and MCC measures. The model trained on all b-values performs best on the b=400 shell according to the AUROC and H-measure scores. However, the models trained on 2 subjects are not capable of adjusting from the lower to highest b-value regime. Hence, by increasing the training data from 2 to 34 subjects, the networks learned a qualitatively different, more general feature extraction.

trained on	tested on	AUROC	H	MCC	FP
34 subjects					
all	0	0.995 [0.994, 0.996]	0.914 [0.905, 0.922]	**0.801** [0.727, 0.859]	12 [10, 14]
0 400 1000	0	**0.996** [0.993, 0.997]	**0.918** [0.895, 0.936]	0.795 [0.761, 0.826]	12 [11, 13]
0	0	0.994 [0.986, 0.997]	0.909 [0.870, 0.938]	0.799 [0.779, 0.818]	12 [11, 12]
all	400	1.000 [0.999, 1.000]	0.995 [0.985, 0.998]	0.924 [0.789, 0.975]	15 [9, 23]
400	400	1.000 [0.999, 1.000]	0.996 [0.988, 0.998]	0.925 [0.687, 0.986]	15 [6, 23]
0 400 1000	400	1.000 [0.999, 1.000]	**0.99653**	**0.942** [0.927, 0.954]	10 [9, 11]
all	1000	**0.9994** [0.9993, 0.9994]	0.966 [0.965, 0.968]	0.901 [0.812, 0.951]	26 [20, 36]
0 400 1000	1000	**0.9994** [0.9994, 0.9995]	**0.967** [0.965, 0.969]	0.919 [0.900, 0.935]	20 [17, 22]
1000	1000	0.9992 [0.9992, 0.9993]	0.966 [0.959, 0.971]	**0.925** [0.840, 0.967]	19 [14, 28]
all	2600	0.994 [0.994, 0.995]	0.947 [0.943, 0.950]	0.783 [0.694, 0.851]	63 [50, 74]
2600	2600	**0.9952** [0.9950, 0.9954]	**0.954** [0.948, 0.959]	**0.810** [0.724, 0.874]	53 [44, 66]
0 400 1000	2600	0.995 [0.992, 0.996]	0.939 [0.923, 0.951]	0.678 [0.570, 0.770]	115 [90, 138]
1000	2600	0.992 [0.988, 0.995]	0.917 [0.879, 0.944]	0.590 [0.180, 0.904]	213 [91, 424]
all	0 400 1000	0.9993 [0.9992, 0.9994]	0.965 [0.963, 0.968]	0.897 [0.798, 0.951]	53 [39, 73]
0 400 1000	0 400 1000	0.9993 [0.9991, 0.9995]	**0.967** [0.964, 0.970]	**0.912** [0.898, 0.925]	42 [38, 45]
all	all	0.9979 [0.9978, 0.9981]	**0.953** [0.948, 0.958]	**0.854** [0.773, 0.910]	116 [89, 139]
0 400 1000	all	**0.9980** [0.9974, 0.9985]	0.949 [0.938, 0.958]	0.814 [0.747, 0.866]	157 [128, 181]
2 subjects					
all	0	0.992 [0.979, 0.997]	0.902 [0.882, 0.919]	0.791 [0.654, 0.883]	13 [9, 16]
0 400 1000	0	**0.993** [0.986, 0.996]	0.897 [0.867, 0.921]	**0.802** [0.704, 0.873]	10 [9, 12]
0	0	**0.993** [0.965, 0.999]	**0.917** [0.716, 0.980]	0.775 [0.724, 0.819]	12 [10, 15]
all	400	**1.000** [0.999, 1.000]	**0.99653**	0.927 [0.858, 0.963]	13 [10, 18]
0 400 1000	400	**1.000** [0.999, 1.000]	0.996 [0.988, 0.998]	**0.948** [0.913, 0.970]	9 [7, 11]
400	400	0.999 [0.999, 1.000]	0.987 [0.976, 0.993]	0.937 [0.784, 0.984]	12 [7, 21]
all	1000	0.9992 [0.9991, 0.9993]	0.965 [0.954, 0.973]	0.892 [0.829, 0.933]	29 [22, 32]
0 400 1000	1000	**0.9993** [0.9993, 0.9994]	**0.970** [0.969, 0.972]	0.907 [0.893, 0.920]	24 [22, 25]
1000	1000	0.9989 [0.9989, 0.9990]	0.961 [0.959, 0.962]	**0.928** [0.884, 0.956]	16 [11, 21]
all	2600	**0.994** [0.994, 0.995]	**0.942** [0.922, 0.956]	0.742 [0.448, 0.911]	88 [44, 131]
2600	2600	**0.994** [0.989, 0.996]	0.935 [0.897, 0.960]	**0.802** [0.577, 0.923]	60 [38, 94]
0 400 1000	2600	0.987 [0.963, 0.996]	0.880 [0.792, 0.934]	0.000 [0.000, 1.000]	836 [85, 1242]
1000	2600	0.985 [0.932, 0.997]	0.865 [0.614, 0.963]	0.000 [0.000, 1.000]	836 [23, 1242]
all	0 400 1000	0.9990 [0.9989, 0.9992]	0.962 [0.959, 0.964]	0.891 [0.847, 0.924]	55 [45, 60]
0 400 1000	0 400 1000	0.9990 [0.9984, 0.9994]	**0.964** [0.952, 0.973]	**0.910** [0.892, 0.925]	43 [40, 47]
all	all	**0.998** [0.997, 0.998]	**0.947** [0.943, 0.951]	**0.832** [0.664, 0.926]	142 [89, 191]
0 400 1000	all	0.993 [0.979, 0.998]	0.907 [0.866, 0.937]	0.536 [0.045, 0.966]	879 [127, 1289]

Table 6.27.: Comparison of classifier performance of *scratch_2233d* models trained on all b-values to models trained on a subset of the b-values. The table is split into two parts, the top half showing data for 34 training subjects, the bottom half for 2 training subjects. Performance analysis is split by b-value using the b-values the respective models were trained on. For models trained on the lower 3 shells results are also shown for the b=2600 shell. All models trained on 34 subjects had at most 2 false negative volumes. The highest false negative number of 9 occurred for a single model trained on 2 subjects' b=1000 data and tested on b=2600. The best average performance values in each test b-value scenario are highlighted in bold, ties are left in light.

6.4.5. Comparison to human inter- and intra-operator variability

Inter- and intra-operator variability can be used to compare the network's performance with human performance. The ground truth labels for all evaluations are defined as the first annotations of the intra-operator comparison, which were generated in the same time-frame and by the same person as the training labels.

Using models that perform best on the test set for comparison with human inter-operator and intra-operator variability would bias the comparison as both data sets contain the same test data. Therefore, an ensemble of models, as a best guess of high performing networks, was defined upfront, prior to training and architecture analysis on the test set. An ensemble of models likely performs equally or better than a randomly chose single model as the sum of weak learners boosts performance if they have different strengths and weaknesses (boosting) and hedges the risk of choosing a poor performing model. The ensemble denoted as *VGGs* consists of two randomly selected models of the architectures *scratch_22*, *scratch_223*, *scratch_2233*, and *vgg16_22333*, all trained on 34 subjects using balanced training. Human performance is compared to that of the ensemble average vote of all 8 models on 25 test augmentations of all slices and orientations.

Averaging the results of multiple slices was a likely candidate for higher consistency [Kelly et al., 2017] but to reiterate, the ensemble was defined prior to knowing the individual model's generalisation performance other than on the small validation set and blinded to the effect of training and test-augmentation strategies other than by proxy to related research [Valle et al., 2017].

b	variability	H	$H^{\vee}$	AP	AUROC	sp95r	MCC	TN	FN	TP	FP
0	intra	1	1	1	1	1	1	67	0	9	0
0	inter	1	1	1	1	1	1	68	0	8	0
0	VGGs at intra	1	1	1	1	1	1	67	0	9	0
0	VGGs at inter	1	1	1	1	1	1	68	0	8	0
400	intra	0.99	0.976	0.964	0.998	1	0.98	216	0	27	1
400	inter	1	1	1	1	1	1	211	0	21	0
400	VGGs at intra	1	1	1	1	1	0.942	214	0	27	3
400	VGGs at inter	1	1	1	1	1	1	211	0	21	0
1000	intra	1	1	1	1	1	1	308	0	33	0
1000	inter	1	1	1	1	1	1	297	0	31	0
1000	VGGs at intra	1	1	1	1	1	1	308	0	33	0
1000	VGGs at inter	1	1	1	1	1	1	297	0	31	0
2600	intra	0.963	0.819	0.742	0.991	0.999	0.854	462	0	23	8
2600	inter	0.963	0.838	0.767	0.992	0.999	0.868	406	0	23	7
2600	VGGs at intra	0.986	0.93	0.949	0.998	0.994	0.788	457	0	23	13
2600	VGGs at inter	0.984	0.93	0.953	0.998	0.993	0.884	407	0	23	6
all	intra	0.982	0.942	0.911	0.996	1	0.95	1053	0	92	9
all	inter	0.985	0.95	0.922	0.996	1	0.957	982	0	83	7
all	VGGs at intra	0.994	0.981	0.995	1	0.997	0.916	1046	0	92	16
all	VGGs at inter	0.994	0.979	0.996	1	0.998	0.963	983	0	83	6

Table 6.28.: Inter-operator and intra-operator variability on the performance of outlier volume detection compared to an ensemble of neural networks. For each variability setting, test volumes that either of the raters deemed ambiguous were disregarded. $M^{\vee}$ stands for the H-measure evaluated on the dual problem of detecting good volumes.

However, human raters had the freedom to rate volumes as ambiguous with the aim of boosting training label quality. This impedes direct comparison of binary classification performance. Best inter- and intra-operator performance is expected on the agreement of volumes that were labelled as 'keep' or 'reject' in each session. Table 6.28 lists the performance evaluated on two subsets of the test data: excluding test samples where any labels of the intra-operator annotations ('VGGs at intra') or where any of the inter-operator labels contained 'borderline' cases ('VGGs at inter'). Note that the overlap in volumes not labelled as borderline is different in the intra- and inter-operator setting. Hence, test set imbalances differ slightly, and performance values in table 6.28, other than AUROC, can not be compared directly across test settings. Also note that human labels are binary, whereas *VGGs* labels are floating point values. As shown in section 5.3, this biases AP values towards higher human performance values.

Network, intra- and inter-operator performance are perfect on the b=0 and b=1000 shell. In the b=400 data, inter-operator variablity and the network assessed on the same data are perfect but the intra-operator performance is lower: one false positive on the human side and 3 false positives on the network's side. The intra-operator test set contains 12 more volumes, hinting at the inclusion of less clear cases than in the inter-operator test set. The highest disagreement occurs on the b=2600 shell but none of the raters or network ensemble have an AUROC below 0.99. The network ensemble yields a higher number of false positives in comparison with the intra-rater performance and exceeds inter-rater performance. Across all b-values, the network ensemble performs better than the level of human inter-operator agreement and achieves an AUROC above 0.999. In other words, across all b-values, the probability that the network ensemble correctly ranks a pair of volumes that both raters or the same rater labelled twice as good and reject is above 99.9%.

To assess the performance on all data, human ratings can retrospectively be transformed into binary labels by assigning borderline cases to either category. Assuming confidence in the pre-processing pipeline, table 6.29 lists results if borderline labels are assigned to the 'keep' category. However, borderline cases are assigned a value of just below 0.5 ($0.5\text{-}10^{-10}$) to retain the information about label rank between reject, borderline and accept. This does not change sp95r, MCC or the confusion table quantities. However, compared to assigning binary labels, using a rank-preserving value can improve AUROC, H and AP values of inter- and intra-rater performance if the borderline labels are assigned consistently.

On all b-values and when assessed across b-values, the network ensemble is more consistent in ranking image quality than the two operators and achieves a higher AUROC than the repeated annotation of the same operator. If thresholded at 0.5, the networks tend to be more conservative (higher false positive rate) than humans. However, despite not been trained on borderline cases, the network ensemble is able to rank images more reliably than human performance.

Model performance is very similar across architectures and model pooling does not show benefits on the small test dataset labelled twice (see table 6.30). However, results on the 4 subjects labelled by two operators suggest that, if calibrated with a – possibly b-value specific – threshold, the network ensemble outperforms human raters. This is

architecture / human	AUROC	AUROC ± 95% CI	H ± 95% CI	MCC ± 95% CI
Incl. borderline cases				
custom*	0.9994			
scratch*	0.9992			
scratch_223*	0.9993			
vgg16*	0.9993			
intra-operator	0.9954			
inter-operator	0.9911			

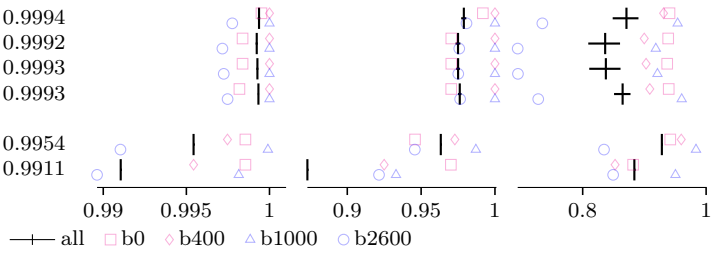


Table 6.30.: Intra and inter-operator performance compared to neural network performances of 25 different models randomly chosen from single or multiple architectures. The test data (4 subjects) uses also borderline cases, which are retrospectively labelled as 'keep': $0.5 - 10^{-10}$. Models are evaluated using 25 augmentations, points and error bars indicate average performance across models.

remarkable as, contrary to human observers, the classifier did not utilise the temporal information, which carries valuable information about motion. Also, they use a simple label averaging procedure to combine labels performed as independent predictions instead of having access to the 3D volume to form a joint decision.

b	variability	H	$H^{\vee}$	AP	AUROC	sp95r	MCC	TN	FN	TP	FP
0	intra	0.946	0.929	0.983	0.999	0.986	0.942	69	1	9	0
0	inter	0.97	0.941	0.982	0.999	0.996	0.882	69	2	8	0
0	VGGs	0.97	0.951	0.991	0.999	0.986	0.947	68	0	10	1
400	intra	0.973	0.952	0.962	0.997	1	0.96	218	1	27	1
400	inter	0.925	0.871	0.944	0.995	0.993	0.853	219	7	21	0
400	VGGs	1	1	1	1	1	0.927	215	0	28	4
1000	intra	0.987	0.98	0.998	1	1	0.984	310	1	33	0
1000	inter	0.933	0.923	0.976	0.998	0.976	0.95	310	3	31	0
1000	VGGs	1	1	1	1	1	0.969	308	0	34	2
2600	intra	0.946	0.793	0.739	0.991	1	0.835	466	1	23	8
2600	inter	0.921	0.802	0.746	0.99	1	0.849	467	1	23	7
2600	VGGs	0.986	0.933	0.95	0.998	0.994	0.761	458	0	24	16
all	intra	0.963	0.914	0.908	0.995	1	0.928	1063	4	92	9
all	inter	0.873	0.831	0.867	0.991	0.965	0.884	1065	13	83	7
all	VGGs	0.983	0.971	0.993	0.999	0.997	0.888	1049	0	96	23

Table 6.29.: Inter-operator and intra-operator variability compared to an ensemble of neural networks. In contrast to table 6.28, all cases that raters deemed borderline labelled as $0.5 - 10^{-10}$, which marks ambiguous volumes as usable volumes in binary classification but preserves the ranking with respect to accept and reject. Note that all test settings share the same data, performance values are therefore directly comparable.

6.5. Conclusions

Neural network architecture search is typically performed manually due to the large search space with non-continuous and conditionally-dependent variables [Elsken, Metzen, Hutter, 2017] and hyper-parameter are selected via cross-validation. However, recent

work showed that the search for optimal network design of vision architectures can be automated by algorithms that are designed to optimise learning [Andrychowicz et al., 2016; Perez, Wang, 2017] or to design performant and efficient network architectures [Zoph et al., 2017]. Given sufficient compute resources (months to years of GPU time [Elsken, Metzen, Hutter, 2017]) and if model performance is the primary goal, then automated architecture search and training optimisation are likely to yield the best results [Zoph, Le, 2016; Real et al., 2017; Liu et al., 2017; Jaderberg et al., 2017; Andrychowicz et al., 2016].

However, by exploring model and data parameters such as data sampling and grouping, it is possible to learn about the structure of the data through training a number of neural networks and analysing how those models perform under these conditions. Contrary to the notion of "black-boxes", once a network is trained - or even during training - it can be used to gain information about the structure of the data. See section 6.7.2 for an exemplary "dissection" of spatial saliency maps of intermediate feature representations and importance maps of areas that contribute the most to the class decision.

More important than the model architecture is the size of the training dataset and the training method. Class imbalance reduces generalisation performance and hinders robust learning. Minority class oversampling improves performance and stabilises training. Cost-weighting decreases performance. Averaging classifications of multiple transformed test images improves model performance and yields more reliable model rankings, helpful for model selection and their application.

An ensemble of networks is perfectly capable of replacing the human annotators and can improve label consistency even when trained on human-generated labels. For the b=2600 shell, an ensemble of networks performs better in ranking image quality (table 6.29) and individual models, especially when trained on the lower 3 shells, yield rates of motion corrupted volumes that are more consistent with the other shells than that of the data they were trained on. This hints at a performance that is above the human level if applied to data with the same characteristics the networks were trained on. The model ensemble was able to cope with a concept drift (functional relation change) from learning to classify volumes that human raters were confident to label, to rating borderline cases at super-human performance.

Progress in classification algorithms due to more sophisticated models might be counteracted by their higher volatility to changing environments that hasn't been taken into account [Hand, 2006], for instance due to the selection of the dataset [Torralba, Efros, 2011]. Hence, generalisation performance has to be tested on other cohorts with data acquired using different sequence parameters and on different scanner hardware. However, the relatively high performance of models trained on a small number of subjects, either from scratch or reusing weights of the pre-trained VGG16 network, suggests that it is possible to transfer the models from the dHCP cohort to a new cohort using a small number of training samples for fine-tuning.

6.6. Appendix: model architectures trained from scratch

Table 6.31.: *scratch_2*

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 1)	0
2	Conv2D 3x3	relu	(64, 99, 8)	80
3	Conv2D 3x3	relu	(64, 99, 8)	584
4	MaxPooling2D 2x2		(32, 49, 8)	0
5	Flatten		(12544,)	0
6	Dense	relu	(8,)	100360
7	Dense	relu	(8,)	72
8	Dense	sigmoid	(2,)	18

Table 6.32.: *scratch_22*

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 1)	0
2	Conv2D 3x3	relu	(64, 99, 8)	80
3	Conv2D 3x3	relu	(64, 99, 8)	584
4	MaxPooling2D 2x2		(32, 49, 8)	0
5	Conv2D 3x3	relu	(32, 49, 16)	1168
6	Conv2D 3x3	relu	(32, 49, 16)	2320
7	MaxPooling2D 2x2		(16, 24, 16)	0
8	Flatten		(6144,)	0
9	Dense	relu	(8,)	49160
10	Dense	relu	(8,)	72
11	Dense	sigmoid	(2,)	18

Table 6.33.: *scratch_223*

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 1)	0
2	Conv2D 3x3	relu	(64, 99, 8)	80
3	Conv2D 3x3	relu	(64, 99, 8)	584
4	MaxPooling2D 2x2		(32, 49, 8)	0
5	Conv2D 3x3	relu	(32, 49, 16)	1168
6	Conv2D 3x3	relu	(32, 49, 16)	2320
7	MaxPooling2D 2x2		(16, 24, 16)	0
8	Conv2D 3x3	relu	(16, 24, 32)	4640
9	Conv2D 3x3	relu	(16, 24, 32)	9248
10	Conv2D 3x3	relu	(16, 24, 32)	9248
11	MaxPooling2D 2x2		(8, 12, 32)	0
12	Flatten		(3072,)	0
13	Dense	relu	(8,)	24584
14	Dense	relu	(8,)	72
15	Dense	sigmoid	(2,)	18

Table 6.34.: *scratch_2233*

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 1)	0
2	Conv2D 3x3	relu	(64, 99, 8)	80
3	Conv2D 3x3	relu	(64, 99, 8)	584
4	MaxPooling2D 2x2		(32, 49, 8)	0
5	Conv2D 3x3	relu	(32, 49, 16)	1168
6	Conv2D 3x3	relu	(32, 49, 16)	2320
7	MaxPooling2D 2x2		(16, 24, 16)	0
8	Conv2D 3x3	relu	(16, 24, 32)	4640
9	Conv2D 3x3	relu	(16, 24, 32)	9248
10	Conv2D 3x3	relu	(16, 24, 32)	9248
11	MaxPooling2D 2x2		(8, 12, 32)	0
12	Conv2D 3x3	relu	(8, 12, 64)	18496
13	Conv2D 3x3	relu	(8, 12, 64)	36928
14	Conv2D 3x3	relu	(8, 12, 64)	36928
15	MaxPooling2D 2x2		(4, 6, 64)	0
16	Flatten		(1536,)	0
17	Dense	relu	(8,)	12296
18	Dense	relu	(8,)	72
19	Dense	sigmoid	(2,)	18

Table 6.35.: *scratch_2233d*

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 1)	0
2	Conv2D 3x3	relu	(64, 99, 8)	80
3	Conv2D 3x3	relu	(64, 99, 8)	584
4	MaxPooling2D 2x2		(32, 49, 8)	0
5	Conv2D 3x3	relu	(32, 49, 16)	1168
6	Conv2D 3x3	relu	(32, 49, 16)	2320
7	MaxPooling2D 2x2		(16, 24, 16)	0
8	Conv2D 3x3	relu	(16, 24, 32)	4640
9	Conv2D 3x3	relu	(16, 24, 32)	9248
10	Conv2D 3x3	relu	(16, 24, 32)	9248
11	MaxPooling2D 2x2		(8, 12, 32)	0
12	Dropout (p=0.2)		(8, 12, 32)	0
13	Conv2D 3x3	relu	(8, 12, 64)	18496
14	Conv2D 3x3	relu	(8, 12, 64)	36928
15	Conv2D 3x3	relu	(8, 12, 64)	36928
16	MaxPooling2D 2x2		(4, 6, 64)	0
17	Dropout (p=0.5)		(4, 6, 64)	0
18	Flatten		(1536,)	0
19	Dense	relu	(8,)	12296
20	Dense	relu	(8,)	72
21	Dropout (p=0.5)		(8,)	0
22	Dense	sigmoid	(2,)	18

Table 6.36.: *scratch_22333d*

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 1)	0
2	Conv2D 3x3	relu	(64, 99, 8)	80
3	Conv2D 3x3	relu	(64, 99, 8)	584
4	MaxPooling2D 2x2		(32, 49, 8)	0
5	Dropout (p=0.2)		(32, 49, 8)	0
6	Conv2D 3x3	relu	(32, 49, 16)	1168
7	Conv2D 3x3	relu	(32, 49, 16)	2320
8	MaxPooling2D 2x2		(16, 24, 16)	0
9	Dropout (p=0.2)		(16, 24, 16)	0
10	Conv2D 3x3	relu	(16, 24, 32)	4640
11	Conv2D 3x3	relu	(16, 24, 32)	9248
12	Conv2D 3x3	relu	(16, 24, 32)	9248
13	MaxPooling2D 2x2		(8, 12, 32)	0
14	Dropout (p=0.2)		(8, 12, 32)	0
15	Conv2D 3x3	relu	(8, 12, 64)	18496
16	Conv2D 3x3	relu	(8, 12, 64)	36928
17	Conv2D 3x3	relu	(8, 12, 64)	36928
18	MaxPooling2D 2x2		(4, 6, 64)	0
19	Dropout (p=0.2)		(4, 6, 64)	0
20	Conv2D 3x3	relu	(4, 6, 128)	73856
21	Conv2D 3x3	relu	(4, 6, 128)	147584
22	Conv2D 3x3	relu	(4, 6, 128)	147584
23	MaxPooling2D 2x2		(2, 3, 128)	0
24	Dropout (p=0.5)		(2, 3, 128)	0
25	Flatten		(768,)	0
26	Dense	relu	(8,)	6152
27	Dense	relu	(8,)	72
28	Dropout (p=0.5)		(8,)	0
29	Dense	sigmoid	(2,)	18

Table 6.37.: *scratch_2233d32*

	name	activation	output shape	parameters
1	InputLayer		(64, 99, 1)	0
2	Conv2D 3x3	relu	(64, 99, 32)	320
3	Conv2D 3x3	relu	(64, 99, 32)	9248
4	MaxPooling2D 2x2		(32, 49, 32)	0
5	Conv2D 3x3	relu	(32, 49, 64)	18496
6	Conv2D 3x3	relu	(32, 49, 64)	36928
7	MaxPooling2D 2x2		(16, 24, 64)	0
8	Conv2D 3x3	relu	(16, 24, 128)	73856
9	Conv2D 3x3	relu	(16, 24, 128)	147584
10	Conv2D 3x3	relu	(16, 24, 128)	147584
11	MaxPooling2D 2x2		(8, 12, 128)	0
12	Dropout (p=0.2)		(8, 12, 128)	0
13	Conv2D 3x3	relu	(8, 12, 256)	295168
14	Conv2D 3x3	relu	(8, 12, 256)	590080
15	Conv2D 3x3	relu	(8, 12, 256)	590080
16	MaxPooling2D 2x2		(4, 6, 256)	0
17	Dropout (p=0.5)		(4, 6, 256)	0
18	Flatten		(6144,)	0
19	Dense	relu	(8,)	49160
20	Dense	relu	(8,)	72
21	Dropout (p=0.5)		(8,)	0
22	Dense	sigmoid	(2,)	18

6.7. Appendix: Looking under the hood of the *scratch_22333d* architecture

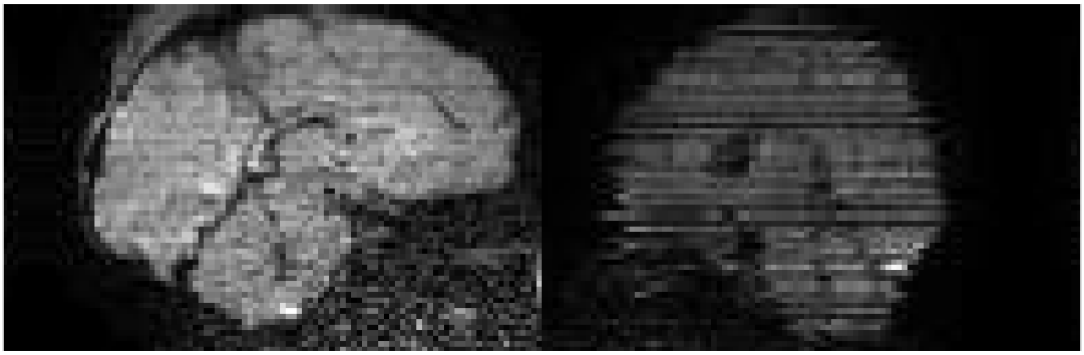
It can be illustrative to understand what aspects of the input data contribute most to a classification algorithm's decision. There are techniques to investigate what affects a general classification algorithm's decision [Baehrens et al., 2010] but recent work on deep neural networks focuses on extracting feature maps using the layer structure of the network or on generating inputs to the network that maximise a neuron's response (activation maximisation). Note that inspecting the activation of single hidden units in a deep neural network is not necessarily informative [Szegedy et al., 2013] as information is spread across units and layers in the network.

Starting from the last layer, one can employ dimensionality reduction techniques such as PCA or t-SNE to visualise a low dimensional representation of the extracted features that are fed to the final classification layer (see [Kelly et al., 2017]). This can give hints about the granularity of the features extracted from the data but becomes less useful for high dimensional outputs (many classes) or intermediate layers. On the input side of the network, one can plot the convolution filters and corresponding weights to get an idea of whether the network contains redundant or "dead" filters. However, human vision and the way deep neural networks perceive images can be very different [Nguyen, Yosinski, Clune, 2014] making interpretation of individual filters or projections of activations tricky. Also, both of those methods are not spatially resolved and provide little information on the network's inner working.

Research on how to visualise features and properties of a neural network in a meaningful way is still in its infancy and often not reproducible [Kindermans et al., 2017]. See [Olah, Mordvintsev, Schubert, 2017] for examples and an introduction into feature visualization in neural networks. Early work in [Zeiler, Fergus, 2013] used a mirrored "deconvolution" network derived from the learned network that allows to trace the contribution to a classification outcome by reverting each convolution, pooling and activation operation in the original network.

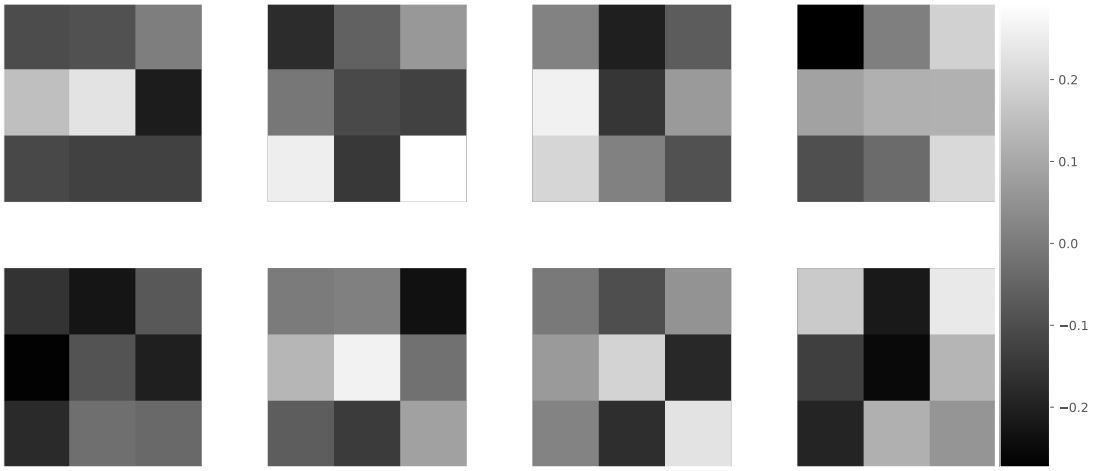
I chose the VGG architecture for its structural simplicity. While the learning process of neural networks is very different from human learning and therefore not particularly intuitive, it is possible to use the trained network to investigate how it processes the image in the early layers. This gives an idea about what kind of low-level features are useful for separating the classes. All images below are from a single network of the 16 layer *scratch_22333d* architecture (see table 6.36).

Figure 6.11.: One good and one rejected sample image of the b=400 shell.



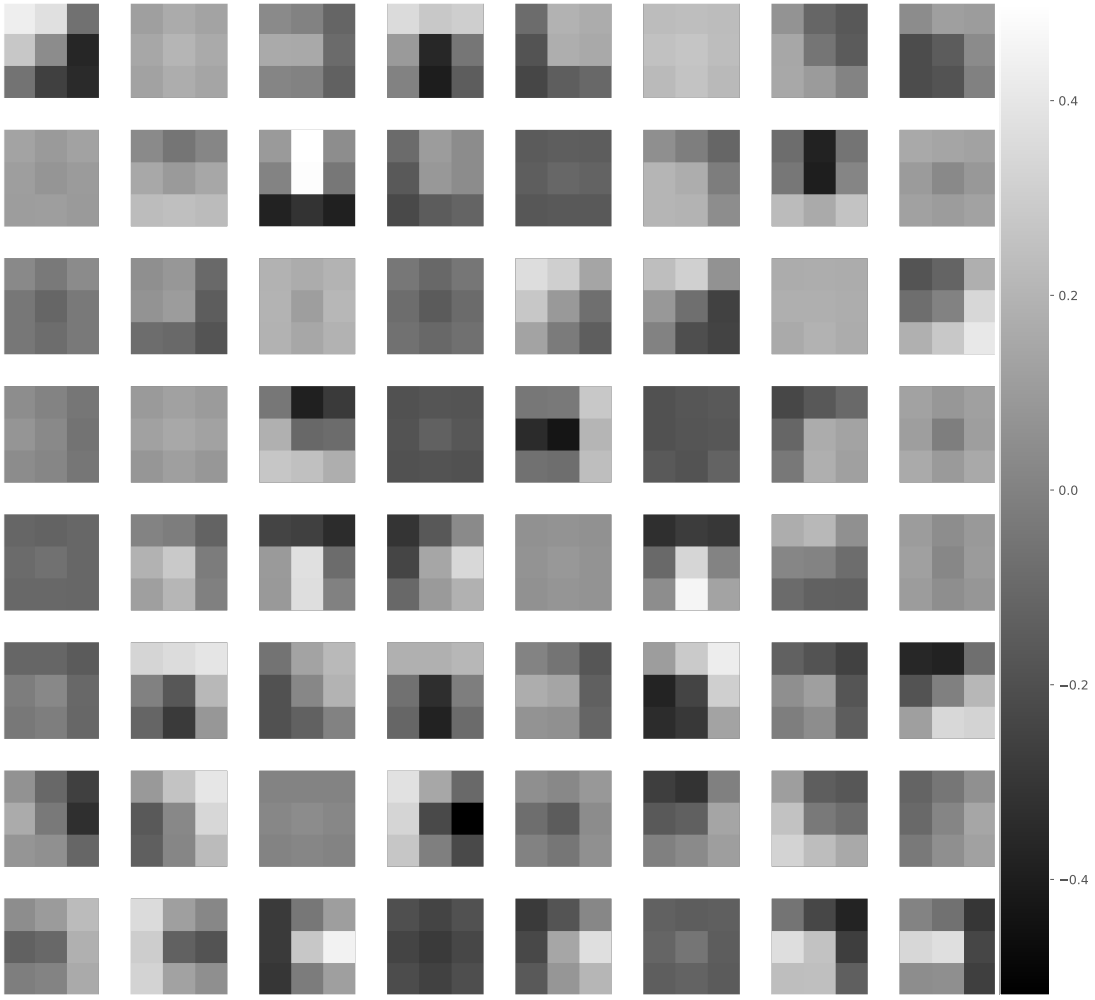
The 8 convolution kernels of the first convolution layer are:

Figure 6.12.: Convolution kernels of the *scratch_22333d* network.



For comparison, the convolution kernels of the original VGG16 model are:

Figure 6.13.: Convolution kernels of the VGG16 network.



6.7.1. Feature representations

The following images display the feature maps that the first convolution and pooling layers of a *scratch_22333d* network produce when they are presented with the acceptable (odd columns) and the rejected (even columns) b=400 images shown above. Early feature maps (see maps after layer 1) resemble edge detector filtered images with different filter directions and strengths. Feature maps of following layers recombine and process these to produce feature maps that seem to highlight horizontal stripes to varying degrees and varying selectivity of their vertical extent (see maps after layer 2), perform background suppression or amplification and extraction of the inferior pial surface. Later layers (see after layer 6) have less clear image characteristics but show a separation of both images in terms of average activation (intensity), which is close to the final goal of using a single threshold to separate both images.

Figure 6.14.: after layer 1

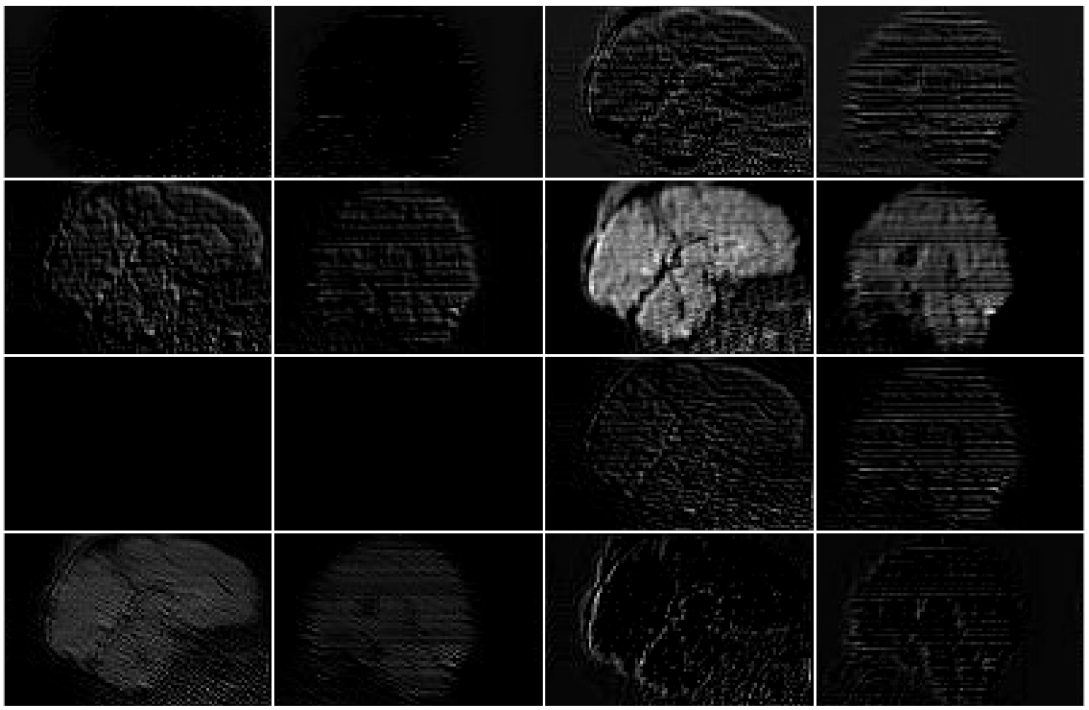


Figure 6.15.: after layer 2

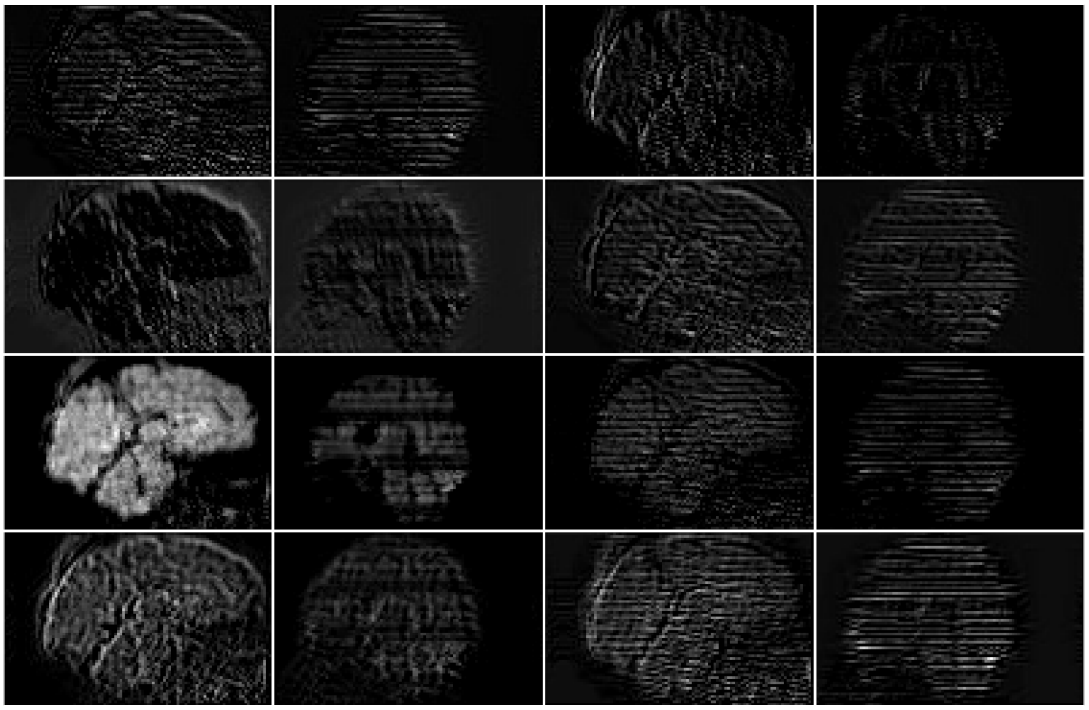


Figure 6.16.: after layer 3

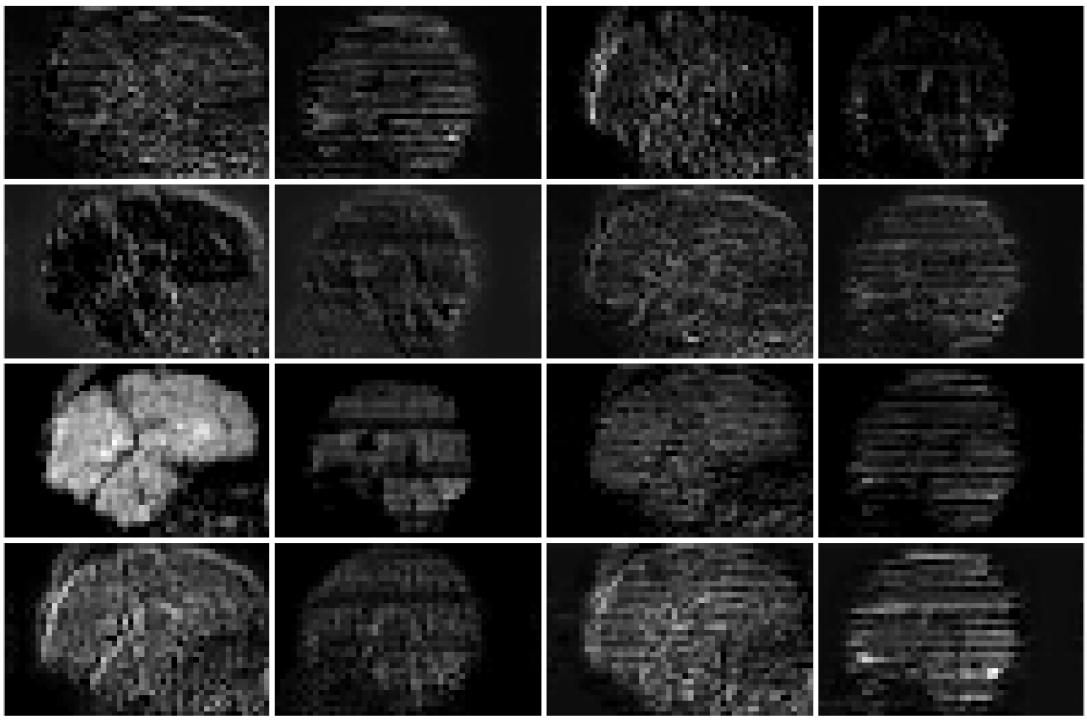


Figure 6.17.: after layer 4

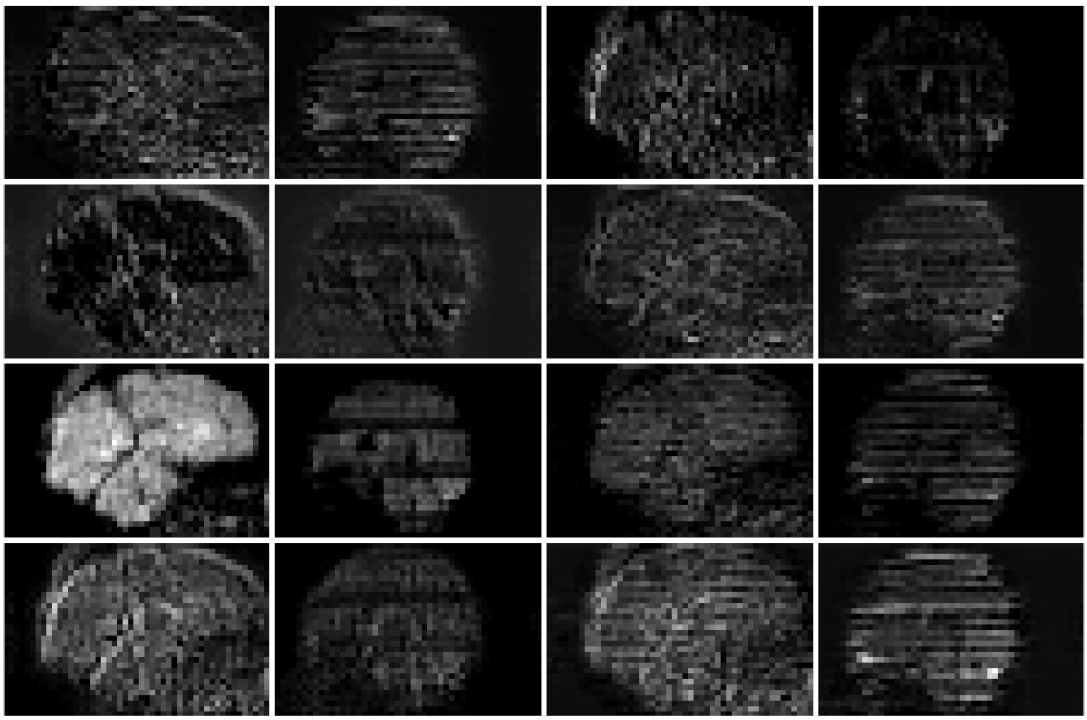


Figure 6.18.: after layer 5

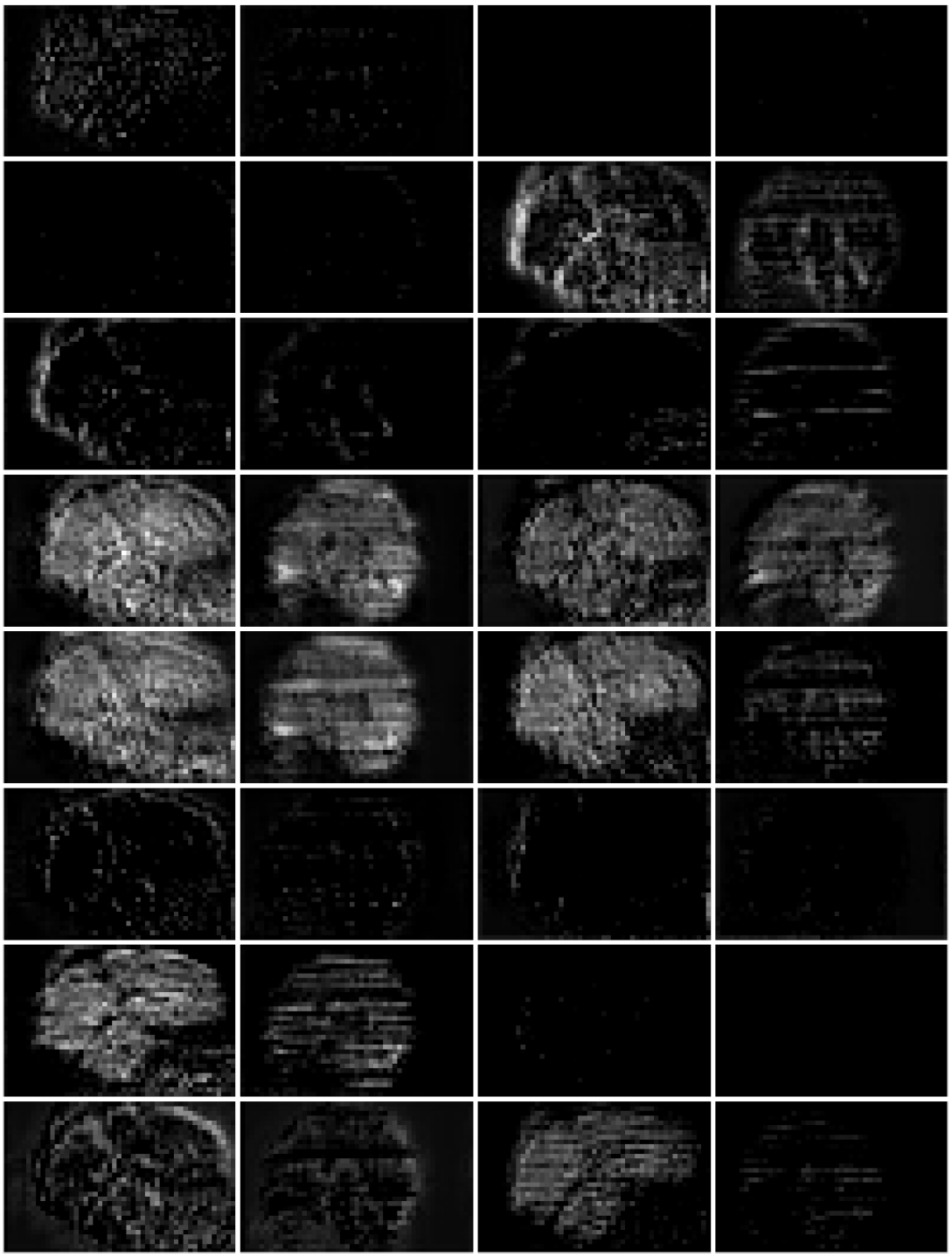
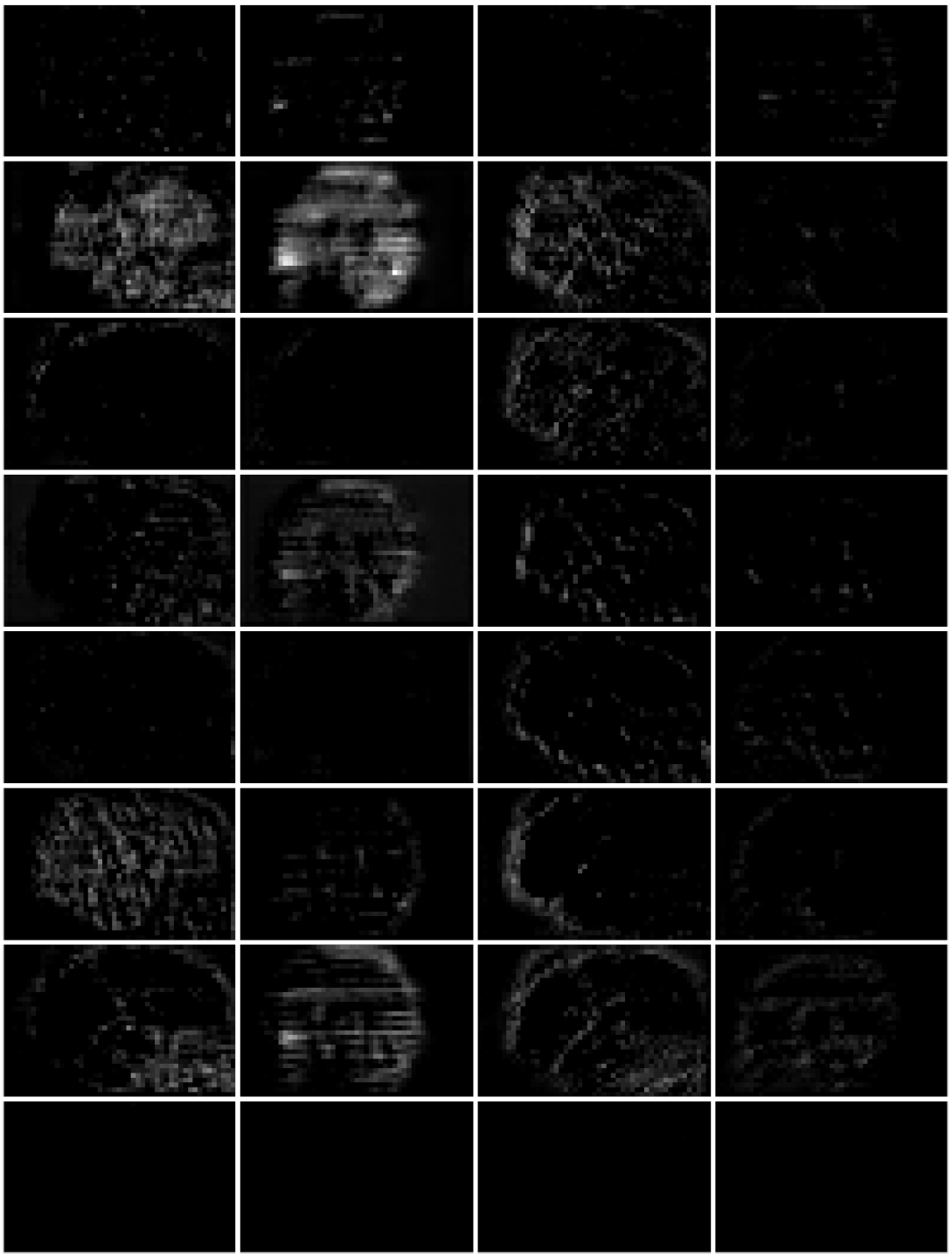


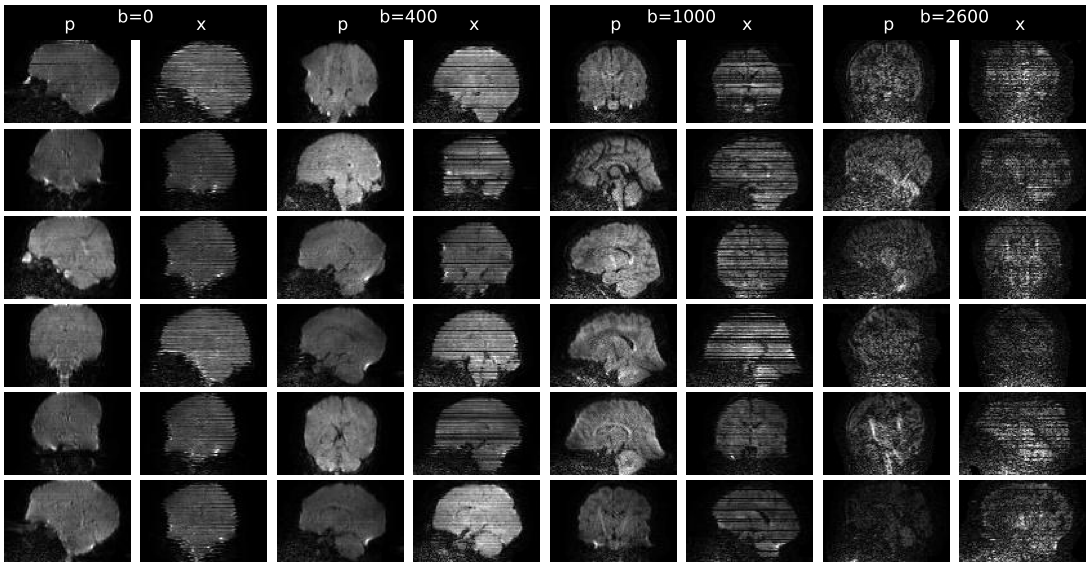
Figure 6.19.: after layer 6



6.7.2. Saliency maps

Furthermore, it is possible to rank input areas by their contribution to the activation of a unit in the network. Creating spatial importance ("saliency") maps [Simonyan, Vedaldi, Zisserman, 2013], requires inverting the information flow (de-convolution, de-pooling and inverting the nonlinearities) and calculating the input layer's Jacobian with respect to the unit's activation. There exist multiple techniques that produce different results. The images below use a method geared towards rectified linear units that propagates only gradients that contribute positively to the activation. The implementation is taken from [Kotikalapudi, contributors, 2017].

Figure 6.20.: The randomly selected input data for generating saliency maps.



The images show pixels that are most important for the activation of 7 units in the last convolution layer of block 2 (after 4 convolution layers) and for the final classification in hot colours. The final layer's sigmoid activation was replaced by a linear function. After the first 4 convolution layers, the networks reacts most strongly to horizontal edges with large contrast. At this level, the network does not seem to place an emphasis on dark bands wider than one or two slices (see fig. 6.21). Also, the units have a degree of overlap in their activations indicating that the network has more capacity at this stage than required. Interestingly, the network seems to have learned to mask the brain images, consistent for different brain positions and some filters focus on the edge of the reconstruction mask fig. 6.23. Mostly for coronal slices, units at this depth react to patterns at the edge of the field of view. This happens at a much lesser degree in sagittal slices. Some filters appear to be less discriminative on the b=0 shell (fig. 6.26) than others (fig. 6.27).

The activations at the very end of the network (fig. 6.28) indicate that the network bases its decision that a volume is acceptable on the majority of the image surface. For outlier volumes, the network reacts most strongly to features at the interface between

brain and background for b=0 images. This changes for higher b-values, where horizontal stripe patterns and "hot spots" located in the brain contribute to a higher degree to the decision.

6.7.2.1. Block 2

Figure 6.21.: saliency map: block 2, filter 1

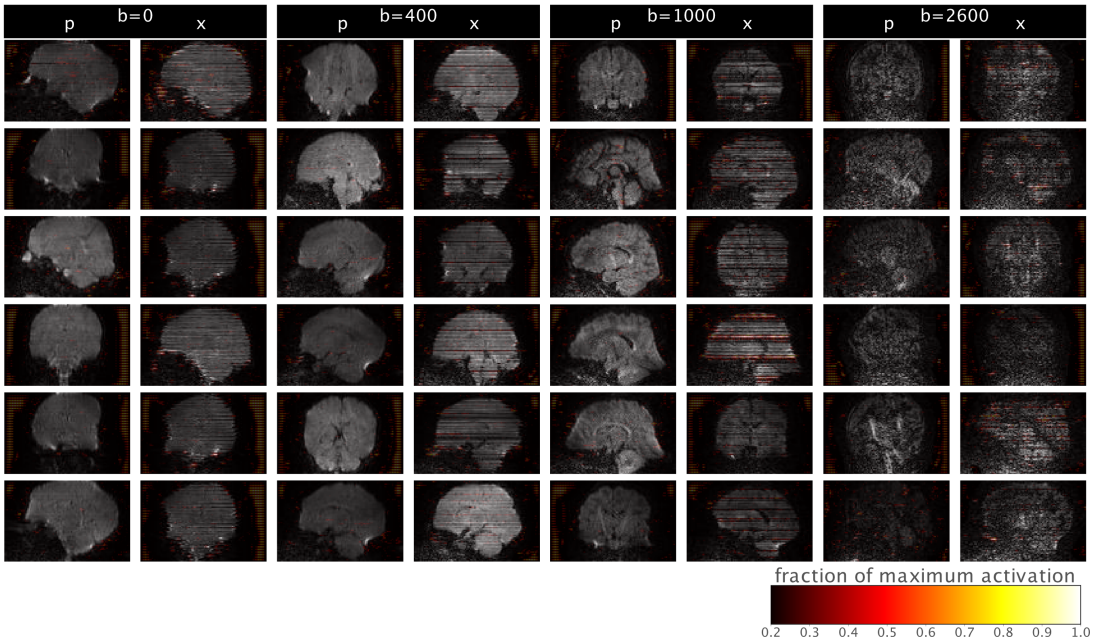


Figure 6.22.: saliency map: block 2, filter 2

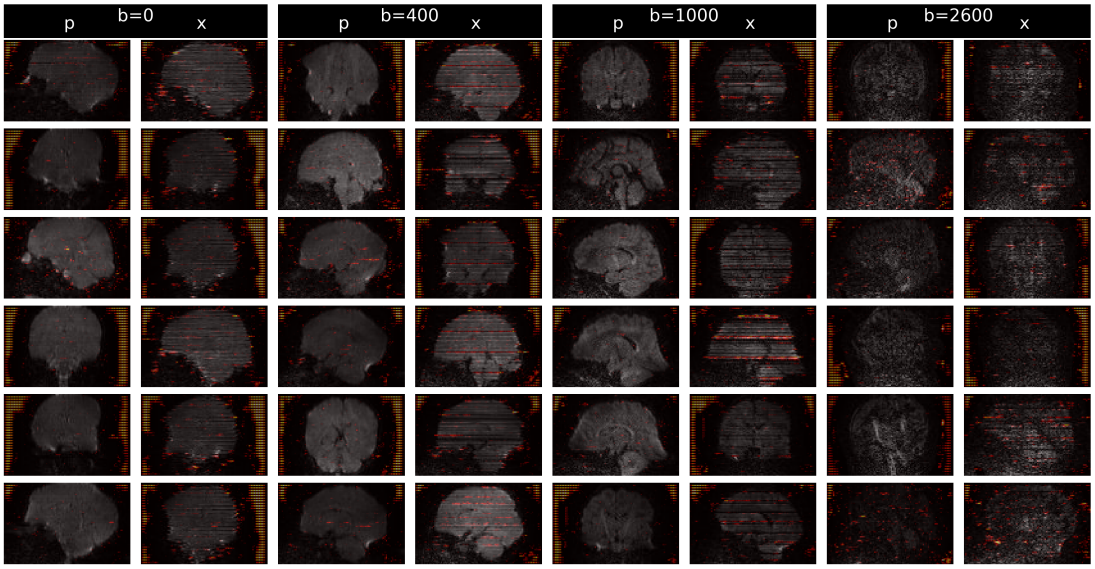


Figure 6.23.: saliency map: block 2, filter 3

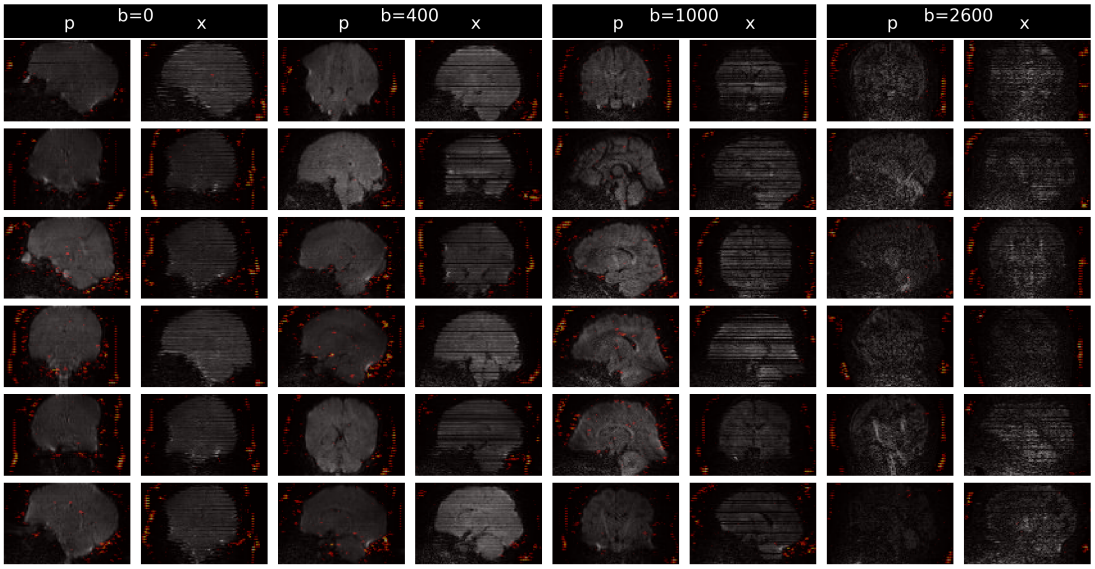


Figure 6.24.: saliency map: block 2, filter 4

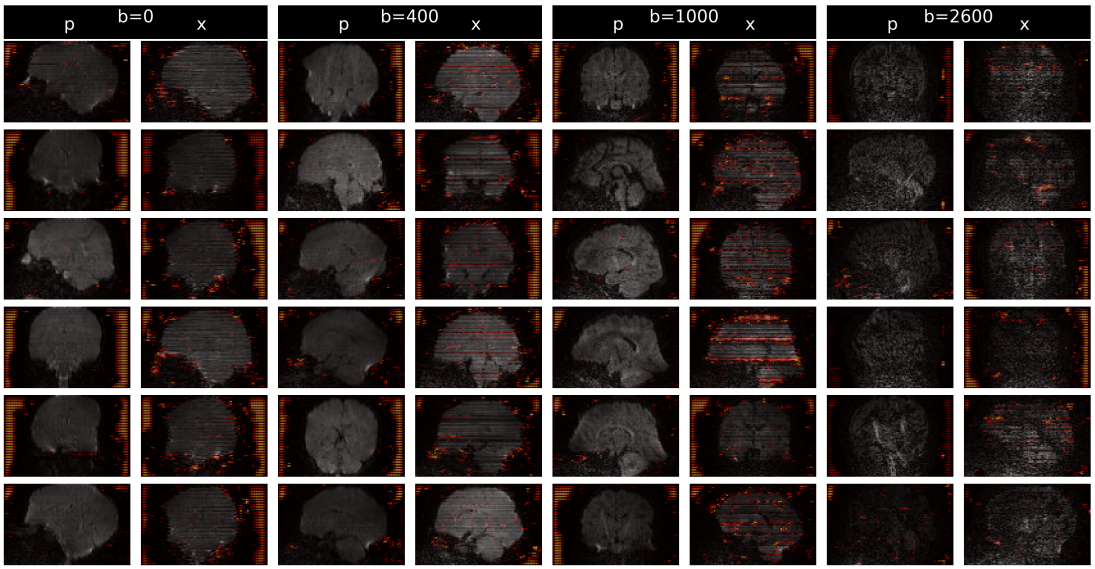


Figure 6.25.: saliency map: block 2, filter 5

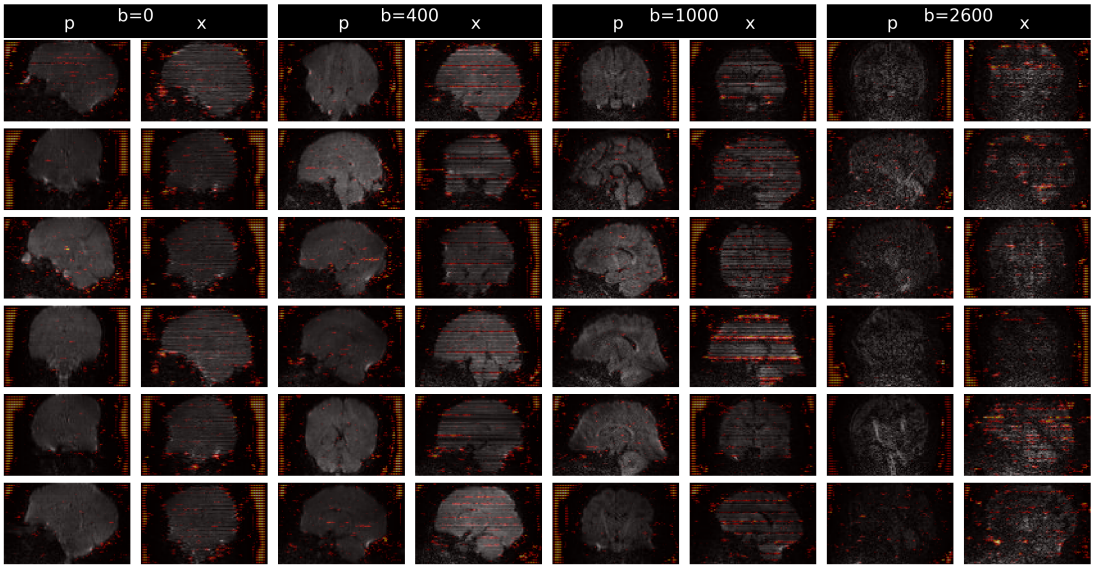


Figure 6.26.: saliency map: block 2, filter 6

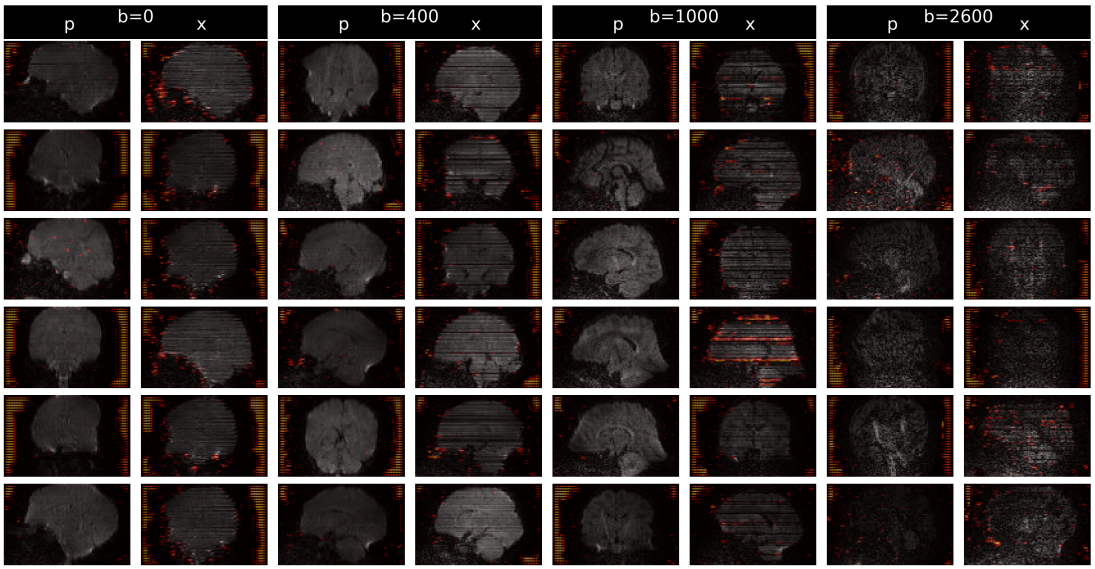
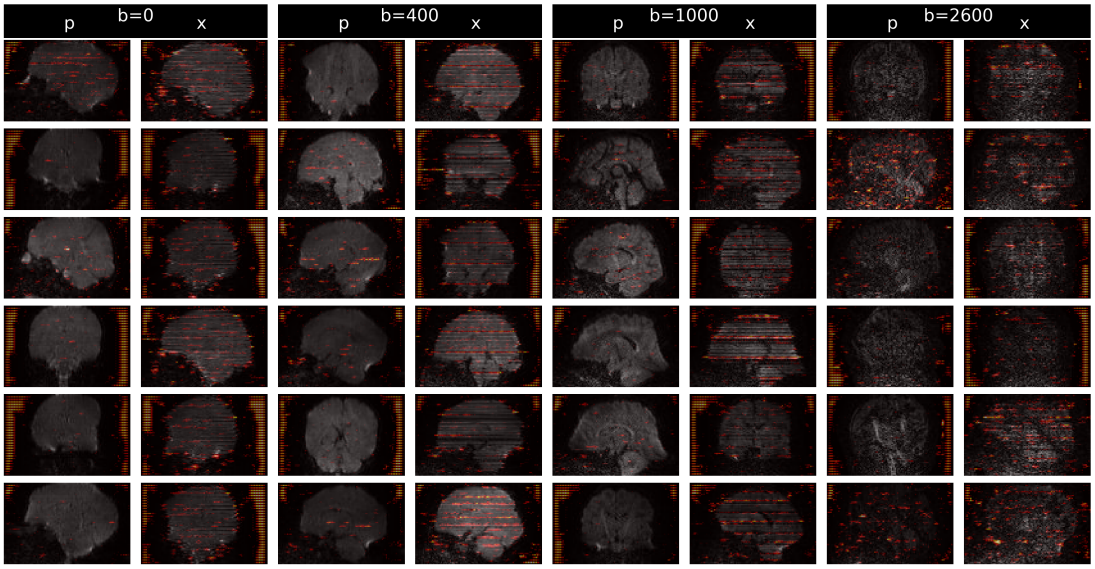
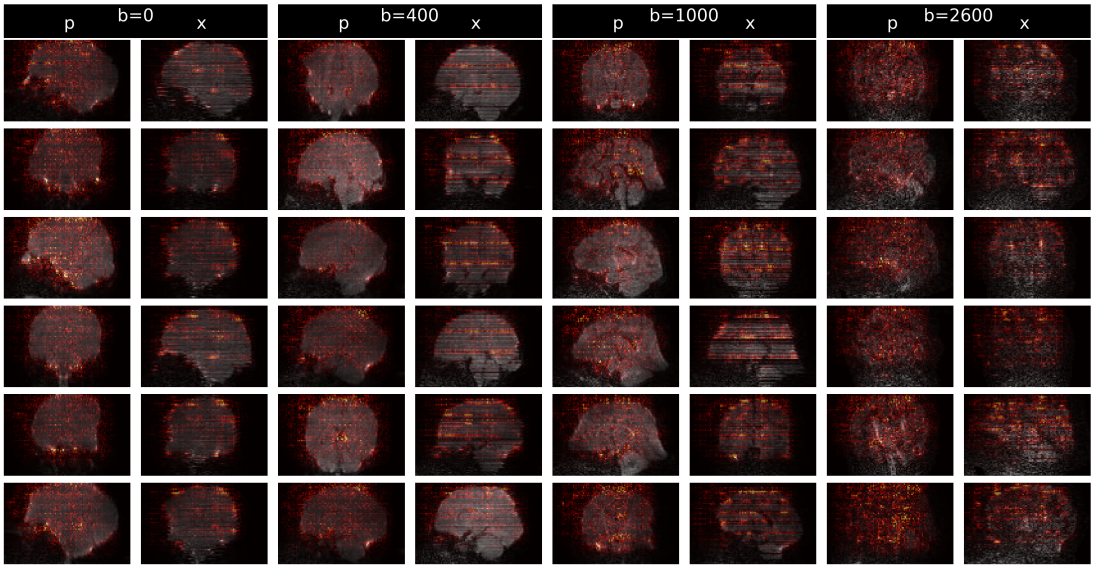


Figure 6.27.: saliency map: block 2, filter 7



6.7.2.2. Final layer

Figure 6.28.: Saliency map of the activation of the last layer.



Chapter 7

Diffusion tensor estimates in the context of changing myelin volume fractions

Contents

7.1. Introduction

In this chapter I perform Monte Carlo simulation experiments of white matter that undergoes a change in myelin content and investigate the effect of changes in relative volume fraction of intra- and extra-axonal space on diffusion tensor indices and fractional anisotropy. This is motivated by the geometrical necessity that a change in myelination of the neonatal brain during development and during demyelination of diseased white

matter is likely to be accompanied by relative changes in tissue compartment volume fractions, and that these may in themselves fully explain the changes observed in these pathologies.

The ratio of the inner axon radius (excluding myelin) to the outer radius including the myelin sheath is called the g-ratio [Rushton, 1951] and changes in the normal developing brain until adulthood [Sherman, Brophy, 2005; Dean III et al., 2016]. For instance, developing human axons in muscle nerves increase in radius up to 5 years of age, while the growth of the myelin sheath surrounding the axons can go on for more than ten years after that [Schröder, Bohl, Bardeleben, 1988]. Decreased g-ratio in cerebral white matter is associated with learning [Fields, 2008] and gives insights into sex-dependent maturation processes during adolescence [Perrin et al., 2009]. At birth, the majority of brain tissue is not myelinated (see fig. 2.1). The cerebellum myelinates early and has a g-ratio of about 0.93 at birth, which falls slowly until adulthood to 0.9 [Dean III et al., 2016]. In the corpus callosum and the corona radiata, the g-ratio drops rapidly from 1.0 at birth to below 0.9 in the first year of life [Dean III et al., 2016].

Starting at 20 year or possibly earlier, the g-ratio increases slowly with age at spatially varying rates [Cercignani et al., 2017]. Besides normal maturation and ageing, pathological changes in g-ratio can occur in the central and peripheral nervous system in the pediatric and adult population. In the whole population in the UK in 2010, 203.4 in 100,000 people were affected by multiple sclerosis [Mackenzie et al., 2014]. The incidence of demyelinating diseases of the central nervous system in children is reported to be between 0.9 and 1.56 per 100,000 [Langer-Gould et al., 2011; Banwell et al., 2009] and that of pediatric-onset multiple sclerosis [Gall et al., 1958; Verhey, Shroff, Banwell, 2013] in particular is estimated to be 0.51 per 100,000 children in America [Langer-Gould et al., 2011]. Measuring changes in g-ratio in-vivo could help distinguishing pathologies such as congenital hypomyelination neuropathy and Dejerine-Sottas syndrome [Balestrini et al., 1991], and it could inform on disease progression in progressive neurodegeneration in amyotrophic lateral sclerosis and primary lateral sclerosis [Kolind et al., 2013]. It has the potential to be a valuable biomarker in patients with chronic multiple sclerosis exhibiting de- and remyelinating lesions, which are associated with disease severity [Albert et al., 2007].

Diffusion tensor imaging is routinely used in clinical and preclinical studies of demyelinating diseases [Aung, Mar, Benzinger, 2013]. Changes in diffusion tensor indices and fractional anisotropy are frequently attributed to different disease and developmental stages and are used as markers of structural white matter integrity [Wheeler-Kingshott et al., 2012; Budde et al., 2009], myelination-related abnormalities [Song et al., 2002] and normal and abnormal brain development [Cheong et al., 2009; Neil et al., 2002; Feldman et al., 2010]. Changes in diffusion tensor quantities are reported to be related to abnormal myelination during white matter maturation [Cheong et al., 2009], demyelination in mouse models [Song et al., 2005a; Song et al., 2003] and multiple sclerosis pathologies [Klawiter et al., 2011], as well as myelin repair [Fox et al., 2011].

However, in general, it is very difficult to unambiguously interpret changes in the diffusion signal and relate them to biological changes in the tissue [Jones, Knösche, Turner, 2013; Jones, Cercignani, 2010; Le Bihan et al., 2006]. The interpretation of

white matter (WM) condition	Fractional Anisotropy (FA)	radial–	axial–	mean–diff.
demyelination	⇓ [1]	⇑ [1,3,4,9]	⇒ [1]	
abnormal myelination /low birth-weight	⇓ [2]	⇑ [2]	⇓ [2]	⇑ [2]
remyelination		⇓ [4,9]		
high myelination / large axons	⇑ [1]	⇓ [1]	⇑ [1]	
multiple sclerosis lesions	⇑ [6]	⇓ [4,6]		⇒ [6] ⇑ [8]
multiple sclerosis normal appearing WM	⇓ [6]		⇓ [6]	⇒ [6]
axonal injury / degeneration	⇓ [1,6]	⇑ [1]	⇓ [1,3,9]	
axonal density	⇑ [0,1]	⇓ [0,1]	⇒ [1]	
WM maturation	⇑ [5,7]	⇒ [7]⇓ [5]	⇑ [7]⇒ [5]	⇒ [5]

Table 7.1.: Reported effects of white matter characteristics on diffusion tensor quantities. Arrows stand for increase (⇑) or decrease (⇓) with the severity of the condition whereas (⇒) stands for no or regionally inconsistent dependency. Sources: 0: [Golabchi et al., 2010], 1: [Feldman et al., 2010], 2: [Cheong et al., 2009], 3: [Song et al., 2003], 4: [Song et al., 2005b], 5: [Bava et al., 2010], 6: [Fox et al., 2011], 7: [Ashtari et al., 2007], 8: [O'Connor et al., 2013], 9: [Sun et al., 2006b]

diffusion tensor-derived measures can be highly confounded in voxels containing non-collinear axon bundles [Wheeler Kingshott, Cercignani, 2009] or voxels contaminated by CSF [Karampinos et al., 2008; Cheng et al., 2011], especially if they are affected by pathology [Mottershead et al., 2003].

Even in the absence of partial volume effects, white matter pathology can cause multiple physiological processes to happen in complex spatial patterns [Burzynska et al., 2010], simultaneously or in succession [Mahad, Trapp, Lassmann, 2015]. Factors that confound the interpretation of diffusion tensor measures include axonal degeneration and associated tissue atrophy [Kim et al., 2007; Ferguson et al., 1997], axonal swelling [Anderson et al., 1996; Fisher et al., 2007] and changes in relative tissue volume fractions and tissue packing density [Golabchi et al., 2010], inflammation [Sun et al., 2006a; Lodygensky et al., 2010; Xie et al., 2010] and oedema [Ebisu et al., 1993]. In particular, oedema caused by different pathophysiological processes such as extracellular oedema with broken (vasogenic oedema) or intact (ionic oedema) blood brain barrier, and oedema affecting cell permeability (cytotoxic oedema), can all exhibit very different diffusion tensor properties.

These processes, all of which or combinations thereof are likely expected in white matter pathologies and abnormal maturation, complicate the interpretation of any observed changes in the diffusion signal. For example, examination of excised nerves in the garfish showed that its non-myelinated olfactory nerve has a higher degree of anisotropy than its myelinated trigeminal and optic nerves [Beaulieu, Allen, 1994]. Yet, most demyelination studies associate reduced myelination with an increase in radial diffusivity (compare table 7.1), with no change in axial diffusivity, which would result in an overall reduction in FA. Clearly, this comparison of unmyelinated and myelinated nerves in the same well-defined environment shows that FA (or AD/RD) alone cannot represent degree of myelination.

Similarly, [Talbott et al., 2016] report that axial and radial diffusivity of severe contusive spinal cord injury in rats, which ablates most axons and myelin sheaths, are virtually

identical to those of axon-sparing chemical demyelination. A further example is found in mouse models used to investigate demyelination in multiple sclerosis lesions, where a decrease in myelin content is associated with an increase in radial diffusivity [Song et al., 2005a; Sun et al., 2006b; Song et al., 2003]. Gadolinium-enhancing MS lesions in humans, however, show the opposite effect on radial diffusivity [Fox et al., 2011]. Finally, [Wang et al., 2011] report no change in radial and axial diffusivity in the centre of the corpus callosum in the cuprizone mouse model of inflammatory demyelination, which they attribute to confounding effects of increased cellularity and vasogenic edema.

Despite these conflicting results, there have been comparatively few studies investigating the effects of changes in neurite tissue density associated with demyelination on the diffusion signal using computer simulations. In this study, I investigate the influence of the degree of myelination on the estimated diffusion tensor in simulations with and without associated neurite density changes. The arrangement of axons and the diffusion simulation are similar to work reported in [Hall, Alexander, 2009]. This study differs from [Hall, Alexander, 2009] in the inclusion of myelin in the substrate and in the way diffusion is modelled within the layers of the myelin sheath. Fieremans et al. investigate diffusion kurtosis changes of parallel axons that undergo demyelination and ablation but in the absence of tissue compaction [Fieremans et al., 2012].

This work aims to specifically investigate the interpretability of diffusion tensor indices in the context of changing g-ratios and to test frequently reported relationships between myelin content and diffusion tensor derived quantities. Therefore, to rule out confounding partial volume effects on the diffusion tensor indices [Vos et al., 2011], the simulated tissue consists of parallel axons in the absence of crossing fibres. This serves as a 'best case' scenario but is an idealised or unrealistic condition, as it disregards non-axonal cell bodies and previous studies found multiple fibre populations in one third of all voxels with an FA above 0.1 [Behrens, Berg, Jbabdi, 2007], and more recently up to 90% of white matter voxels [Jeurissen et al., 2013] at typical in-vivo resolutions. Furthermore, any other potential disease-specific changes in tissue properties are disregarded. Nonetheless, these simulations provide an alternative perspective on factors that might affect DTI-derived measures in pathology.

7.2. Model-based simulation of diffusion

7.2.1. Monte Carlo diffusion simulation

Diffusion from an atomistic point of view can be understood as particles undergoing collisions with other particles and barriers (see section 3.1). While it is impossible to describe the movement of each particle, one can make certain statements about the whole ensemble of particles. Diffusion is simulated as many particles that move in a specified way so that their ensemble average resembles Gaussian diffusion. In particular, the particles have no preferred direction (unbiased) and no history (uncorrelated) but obey the physical boundaries of the simulated tissue. The ensemble average particle displacement for free diffusion after a time period t is $\sqrt{(2d)Dt}$ where d is the dimensionality of the system and D the Gaussian diffusivity. A consequence of the central limit theorem is

that for a sufficient large number of time steps, one can use a fixed step width for all walkers as an approximation for Gaussian diffusion [Hall, Alexander, 2009].

From a statistical physics point of view, diffusion MRI measures the probability distribution of all particles' displacements. This distribution is Gaussian in the case of free diffusion but in general unknown for hindered or restricted diffusion. By simulating the movement of many particles that experience the tissue's diffusive properties, one can estimate the actual probability distribution of this anomalous diffusion without the need to explicitly derive it. This technique is called Markov chain Monte Carlo and is used to simulate various biological processes [Codling, Plank, Benhamou, 2008].

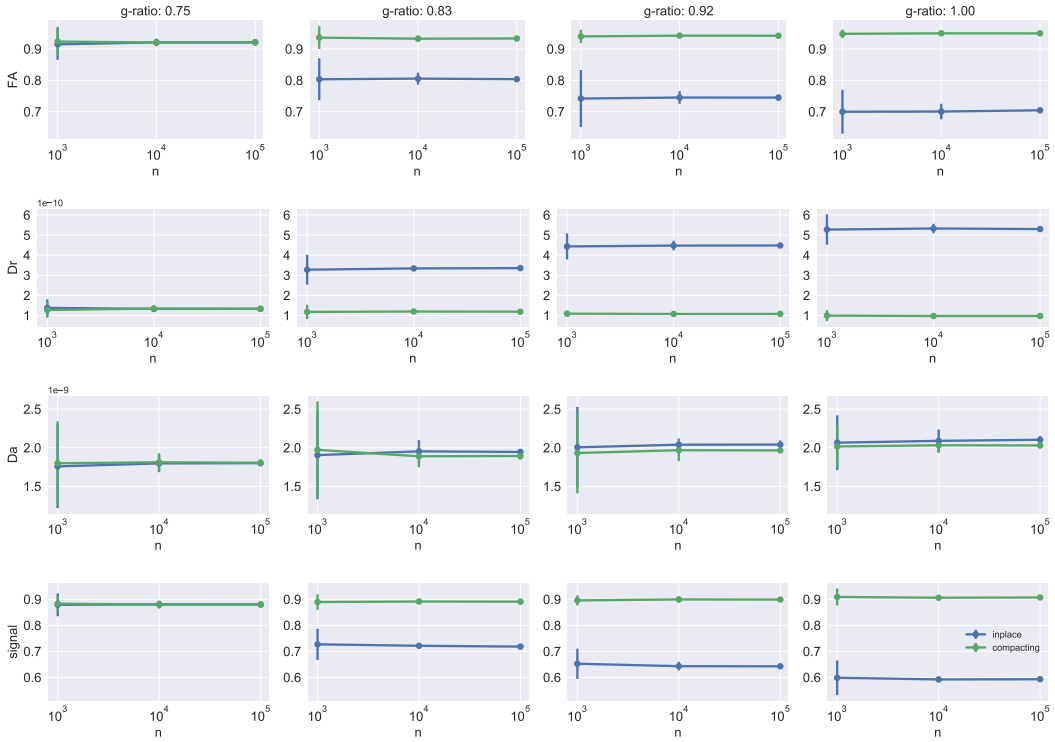


Figure 7.1.: Calibration of the number of walkers n with t = 50,000 time steps on the substrate with the highest axonal packing density (80.6%).

The program datasynth [Panagiotaki et al., 2012] from the camino toolkit [Cook et al., 2006] allows simulating diffusing water molecules in tissue substrates and measure the spin phase during an MRI experiment. In this study, the signal was measured in 61 directions with b-values 1000s/mm^2 and 3000s/mm^2 with a Stejskal-Tanner sequence (see section 3.2.1). The sequence parameters were chosen to lie in the range of achievable parameters for typical clinical scanners: TE=85ms, δ=22.6ms and Δ=32.6ms for b=1000s/mm^2 and TE=105ms, δ=43.2ms and Δ=44.2ms for b=3000s/mm^2. The gradient strength is 33mT/m for both b-values.

The number of particles (N) and time steps (t) used in the MC simulation needs to be calibrated to the substrate to trade off precision and accuracy of the simulations [Hall,

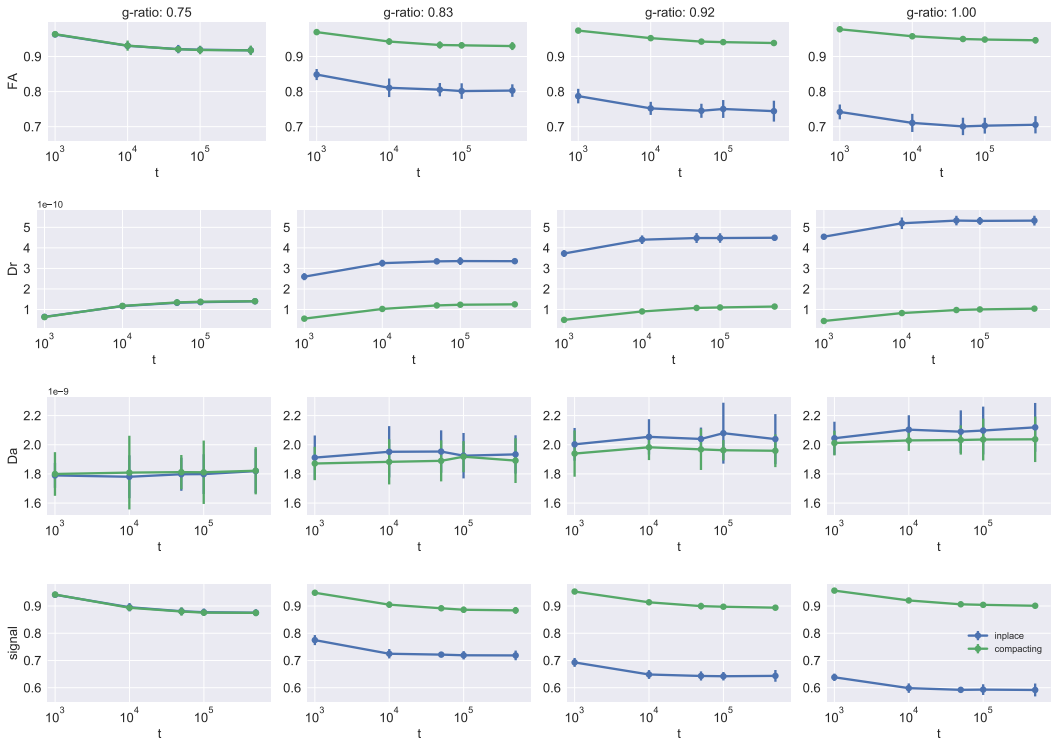


Figure 7.2.: Calibration of the step size t with n = 10,000 walkers on the substrate with the highest axonal packing density (80.6%).

Alexander, 2009]. To calibrate the parameters, a parameter search for N up to 100,000 and t up to 500,000 was run on a substrate with maximum packing density and each parameter combination was repeated ten times to assess the rerun variability. fig. 7.1 and fig. 7.2 show the mean effect of demyelination across reruns on all considered variables for increasing N and t. Based on these plots I deem N=10,000 and t=50,000 to be sufficient to avoid bias while still being computationally feasible. Simulations of the substrate with the highest packing density when fully demyelinated show that with t=50,000 at most 19.6% of the walkers scatter off an axonal membrane at any time step, which in other studies is generally accepted as sufficient temporal resolution for unbiased results [Landman et al., 2010]. An illustration of the walkers that interact with a boundary in a single time step is shown in fig. 7.8 for a subset of this high density substrate.

7.2.2. Modelling white matter

Axons Axons are modelled as double-walled impermeable cylinders. The inner cylinder defines the intra-axonal space in which water can move in any direction until it hits the inner axonal wall, from which it deflects elastically. Myelin was modelled as infinitely thin non-permeable cylindrical layers wrapped concentrically around the inner axon, with walkers restricted to diffusing on the surface of their corresponding myelin layer. In other

words, exchange between myelin layers was assumed to be negligible.

The simulations start at the lowest g-ratio of 0.75 which is derived from data in [Liewald et al., 2014] who report g-ratios of healthy white matter ranging from 0.52 to 0.86 in the superior longitudinal fascicle of a macaque monkey with a median value of 0.74.

Axon arrangement Analogous to [Hall, Alexander, 2009], white matter bundles are modelled as parallel axons with random diameters and randomly organised. The cross section through a unit rectangular cuboid of axons is depicted in fig. 7.3. Axons intersecting with the edges of this cross section are wrapped around on the opposing side to counteract edge effects. The substrate is replicated in the cross-sectional plane so that walkers, which are initially located at random positions within the cuboid, do not diffuse into empty space.

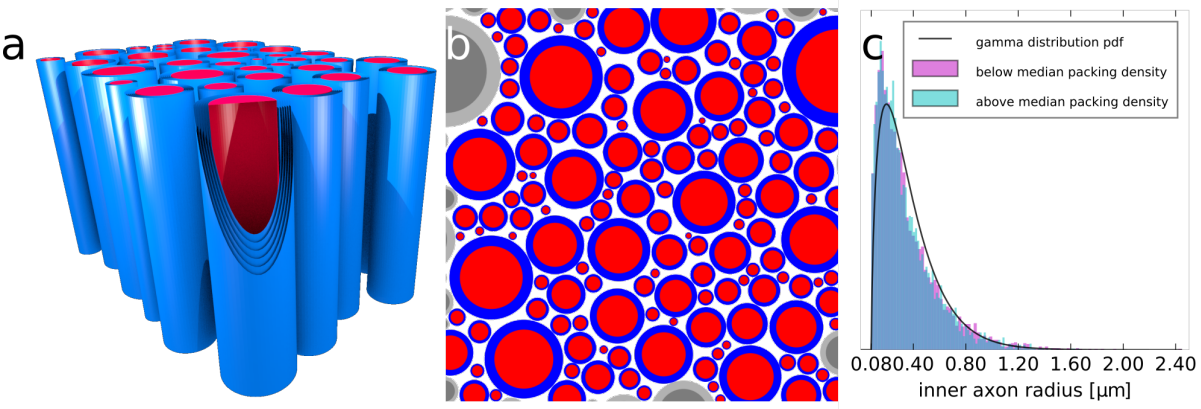


Figure 7.3.: Figures (a) and (b) are visualisations of the simulated substrate showing parallel cylinders with layers of myelin (blue) surrounding the intra-axonal compartment (red). The periodicity of the substrate is illustrated in (b) as grey axon fragments that correspond to axons on the opposing side of the substrate. Figure (c) shows the distributions of simulated axon radii across all substrates split into two groups based on their packing density. Cyan (magenta) bars correspond to substrates with packing densities larger (smaller) than the median packing density of all substrates, overlap is shown in sky blue. The line plot is the probability density of the shifted gamma distribution we fitted to data taken from [Liewald et al., 2014].

Similarly to the AxCaliber model [Yaniv Assaf et al., 2004; Assaf et al., 2008] and simulations in [Hall, Alexander, 2009], the probability of an axon having a specific inner radius r is assumed to follow a gamma distribution whose probability density is defined as

$$p(r;p,s) = \frac{r^{p-1}e^{-\frac{r}{s}}}{s^p\Gamma(p)} \quad \text{for } r > 0 \text{ and } p, s > 0. \tag{7.1}$$

with parameters p (“shape”) and s (“scale”) and Γ the gamma function.

To simulate realistic axon radii, this function was fitted to axon diameter measurement histograms taken from an electron microscopy study of the human occipitofrontal fascicle [Liewald et al., 2014]. Before fitting, the radius distribution was shifted by the minimum observed radius 0.08μm as axons must have a minimum non-zero radius to be biologically plausible. The fitted probability density function has parameters $s = 0.17\mu m$ and $p = 1.68$ and is shown in fig. 7.3.

The non-overlapping placement of axons with relatively high packing density was achieved by brute-force placement of axons starting with the largest ones similar to [Hall, Alexander, 2009]. However, this ordering by size necessitates defining a fixed number of radii that are drawn from the gamma distribution. The constraint to place all axons in the defined volume and the limitation to a small number of axons (170) for efficiency reasons might skew the actual distribution of successfully drawn radii. The minor discrepancy in probability density shape between the assumed gamma distribution and the resulting axon distribution shown in fig. 7.3 (c) is likely due to these reasons. Thankfully, this effect seems independent of packing density, which is defined as the proportion of the cross-sectional area taken up by the myelinated axons in the substrate. The simulated axon packing density was varied from 0.54 to 0.81, which is respectively below and slightly above reported packing densities of 0.80 in the rat brain [Syková, Nicholson, 2008].

Tissue properties In the simulations, the diffusive properties of myelin are modelled differently compared to the rest of the tissue. This is motivated by different water volume fractions inside axons and between myelin sheaths and reported differences in diffusivity and T_2 relaxation time, whereas intra- and extra-axonal tissue are assumed to share these parameters [Akhondi-Asl et al., 2015; Björk et al., 2016]. While myelin-bound protons are ignored as they can not be measured due to their very short T_2 on the order of 10μs [Samsonov et al., 2012], we used T_2=26ms for water trapped between myelin layers, and T_2=80ms elsewhere [Hurley, Mossahebi, 2010; Laule et al., 2004]. [Laule et al., 2004] report 0.369g H_2O per gram myelin and 0.82g H_2O per gram of non-myelinated tissue in normal white matter. Hence, we use free water volume fractions of 0.37 for myelin and 0.82 for non-myelin in the simulations.

ADC values for both environments are difficult to translate into free diffusion lengths as the ADC is also influenced by tissue geometry. A diffusion constant of $2.0\mu s^2/ms$ was used for non-myelinated tissue and $0.5\mu s^2/ms$ for myelin. The latter is based on apparent diffusion constants measurements in frogs by [Andrews, Osborne, Does, 2006], who report the ADC in myelin at room temperature compared to the intracellular compartments being reduced by a factor of 3.1, measured along, and 5.1, if measured perpendicular to the axon. Step lengths for myelin are calculated according to the expected mean squared displacement for 2D Brownian motion as these walkers are effectively confined to a 2D surface. To increase the sampling density of myelin, walkers were placed with equal density in all compartments and the signals were weighted by their respective volume water fraction and T_2 attenuations.

Approximating dispersion Axons in white matter bundles are not perfectly parallel at the resolution level of typical diffusion MRI. The mean intravoxel dispersion in the corpus callosum of humans is 14.6 ± 3.6 degrees [Budde, Annese, 2013], 14.4 degrees in rats [Leergaard et al., 2010] and 12 degrees in owl monkeys [Choe et al., 2012].

Using parallel cylinders, the simulated model can not accurately account for the effect of dispersion on the extracellular space. Therefore the simulated effect of dispersion is limited to a blurring of the signal in the angular domain. This replaces modelling dispersion within a single substrate with averaging the signal of a large number of rotated substrates.

To speed up computation, the signal of the existing Monte Carlo calibration simulations were reused but each was assigned a retrospectively rotated version of the diffusion gradient directions. For each sample the gradient directions were jointly rotated around a random rotation axis by an angle drawn from a normal distribution with standard deviation of 14 degrees.

This was repeated 1000 times and the tensor fit for each sample was performed on the resulting concatenated signal and gradient table.

Rotating gradient directions and concatenating signal vectors has the further advantage that it avoids interpolation artefacts when the signal of those rotated substrates are mapped to the original diffusion gradient directions.

This is equivalent to averaging the signal of 1000 rotated exact replicates of the substrate that do not interact with each other. This is therefore an approximation because, unless dispersion happens at a much coarser spatial resolution than the axonal level, one would expect the signal to be affected by the difference in geometry that dispersion must introduce. Note that all simulations and results are without dispersion unless stated otherwise.

7.2.3. Modelling demyelination

Demyelination in the following simulations is defined as increasing g-ratio that ranges from 0.75, representing healthy mature white matter, to 1 for completely demyelinated axons. This change in myelin content, however, does not define how the remaining tissue behaves. Depending on the process that is causing the change in g-ratio and its speed, one can imagine two different scenarios, which are illustrated in fig. 7.4: myelin either trades space with the extracellular matrix with the axons remaining stationary; or the extracellular volume fraction stays constant during changes in g-ratio.

For demyelination, the latter causes white matter tracts to compact down during demyelination due to the reduced myelin volume fraction. The compacting case is simulated as an increase in the inner radius, exactly compensated by a global down-scaling of the substrate. Despite constant inner axonal radii during demyelination, the decreased size of the substrate in effect increases the intra-axonal volume fraction at the expense of the myelin volume fraction. In the “in place” case, the extracellular space is obviously less hindered. Although the extra-axonal volume fraction remains constant in the compacted case, the number of “pores” per unit area in the extracellular space increases with demyelination causing a decreased average pore size, which in turn results in a more

hindered extracellular matrix.

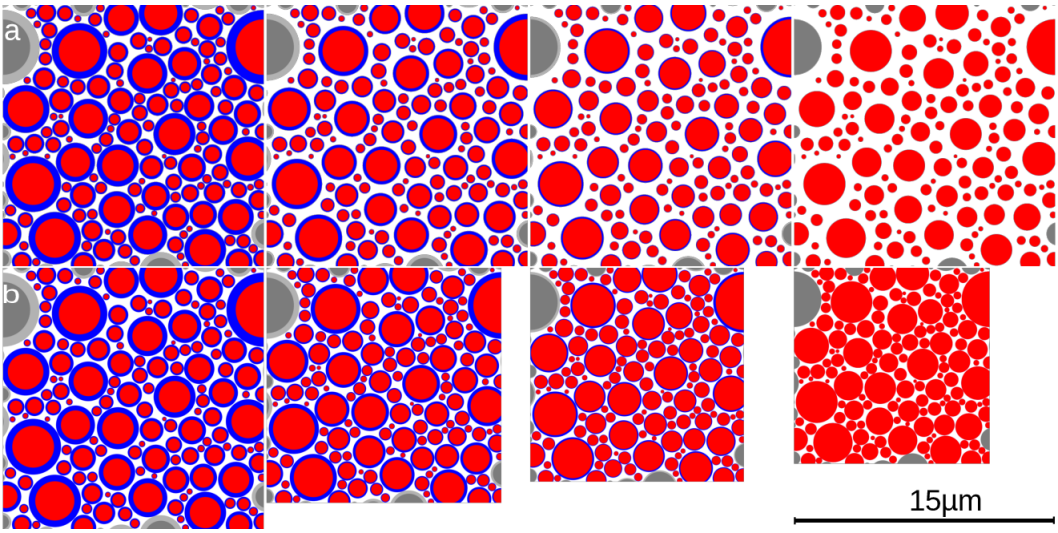


Figure 7.4.: Illustration of simulated demyelination of a substrate where axons, depicted in red, undergo demyelination with linearly increasing g-ratio from left to right. The volume fraction of myelin (blue) is either taken up by extracellular space where axons stay in place (a) or by the intracellular space where the tissue compacts down (b).

7.3. Results

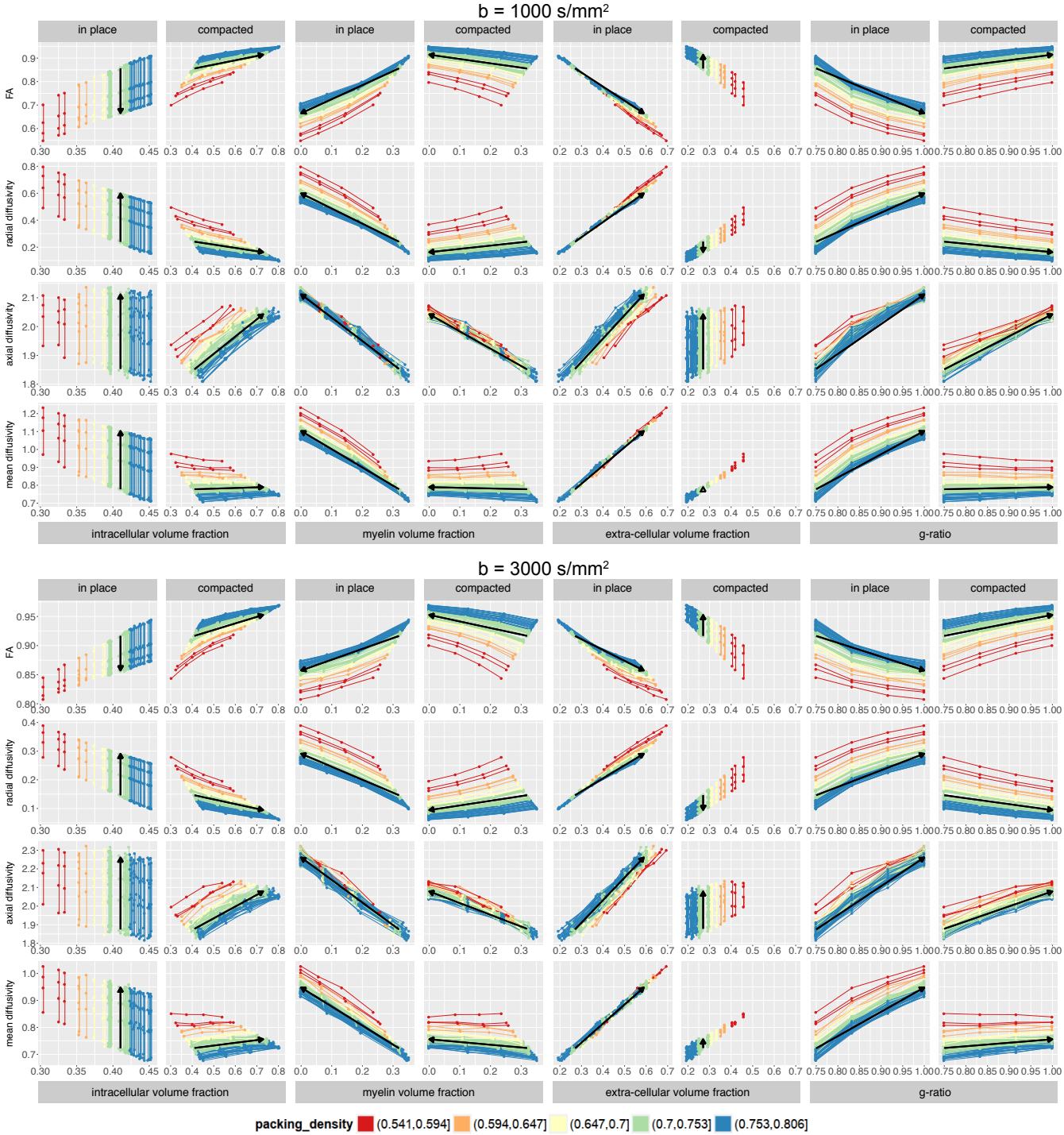


Figure 7.5.: Diffusion tensor measures plotted in relation to changes in intracellular, myelin, and extracellular volume fractions and g-ratio separately for the two demyelination scenarios at b=1000s/mm^2 (top) and b=3000s/mm^2 (bottom). Note that intracellular, extracellular, and myelin volume fractions as well as packing densities are not independent and their relations are determined by the simulation scenario. Arrows represent the direction of change from fully myelinated to demyelinated axons. Diffusivities in units of 10^{-3}mm^2/s. Lines link across different g-ratios for the same substrate and colours denote the initial substrate packing density.

7.3.1. Axial diffusivity

There is a consistent inverse relationship between axial diffusivity and myelin content, irrespective of the demyelination scheme and b-value (fig. 7.5). The difference in axial diffusivity between fully myelinated and fully demyelinated tissue is on the order of 9%, which is mainly driven by the reduced contribution of the myelin signal whose lower mean diffusivity reduces the overall axial diffusivity. In fact, if the myelin signal is not taken into account, the axial diffusivity is independent of the g-ratio with the exception of near-complete demyelination where the diffusion tensor can not represent the stick-like shape of the attenuation profile and the fit overestimates the axial diffusivity.

7.3.2. Radial diffusivity

For both b-values, radial diffusivity increases with demyelination for the "in place" scenario and decreases for the "compacted" case and is therefore not specific to myelin content. In the simulations, it is strongly dependent on the extracellular volume fraction and confounded by the initial tissue packing density, irrespective of the simulated demyelination scenario. In other words, a change in radial diffusivity cannot be attributed to a change in myelin volume fraction if the intra-axonal and extracellular volume fractions are unknown.

7.3.3. FA and mean diffusivity

Fractional anisotropy decreases for substrates where axons stay in place during demyelination and increases for the "compacted" scenario. Axial diffusivity is mostly independent of the demyelination scheme; the FA behaviour can therefore be mainly attributed to the changes in radial diffusivity. Mean diffusivity increases with demyelination in the "in-place" scenario, closely following the increase in extracellular volume fraction. Yet, it is near constant in the "compacted" case where mean diffusivity is positively or negatively correlated with myelin volume fraction depending on the packing density. These findings are valid for both simulated b-values.

7.3.4. Myelin tissue properties

To investigate the effect of the choice of myelin tissue properties, all simulations were repeated for the two extreme cases: (i) ignoring the myelin signal; and (ii) setting all tissue properties (T_2, free water density, diffusivity) to those of non-myelin tissue. For b=1000s/mm^2, all results as stated above are qualitatively still valid. The only qualitative difference in radial diffusivity and FA can be observed in the (ii) scenario in the curves for b=3000s/mm^2 in the "in place" scenario in which FA first decreases and then increases with progressing demyelination (data not shown).

7.3.5. Packing density

All lines in fig. 7.5 correspond to a specific substrate undergoing demyelination. The colour-coding indicates the respective packing density (defined as intra-axonal plus myelin

volume fraction or equivalently 1 - extracellular volume fraction) of this substrate before demyelination. For the compaction scenario this initial packing density, by design, remains constant during demyelination but the range of simulated packing densities allows its effect to be observed. All observed trends are largely independent of initial packing density, yet the relative changes of FA and radial diffusivity are larger for higher packing densities.

7.3.6. Myelin content as a function of AD and RD

Depending on the simulation scenario, a reduction of the myelin volume fraction causes an increase of the relative volume fraction of the extracellular compartment ("in-place") or of the intra-axonal compartment ("compacting") volume fraction. The axonal packing density stays constant in the "compacting" case but decreases in the "in-place" simulations. Therefore, determining any of those volume fractions' influence on the axial or radial diffusivity requires taking those correlations into account.

Partial least squares regression [Abdi, 2010] between all volume fractions and radial and axial diffusivity showed that the intra-axonal volume fraction is the volume fraction that, taken by itself, is the least correlated with any diffusion tensor measure and that myelin- and extracellular volume fractions have a near orthogonal effect on the measured diffusion tensor indices. Hence, in the simulation results, radial and axial diffusivity together capture most of the information about the geometry of the substrate and can therefore be used in a multivariate linear model to predict volume fractions of any tissue compartment (fig. 7.6).

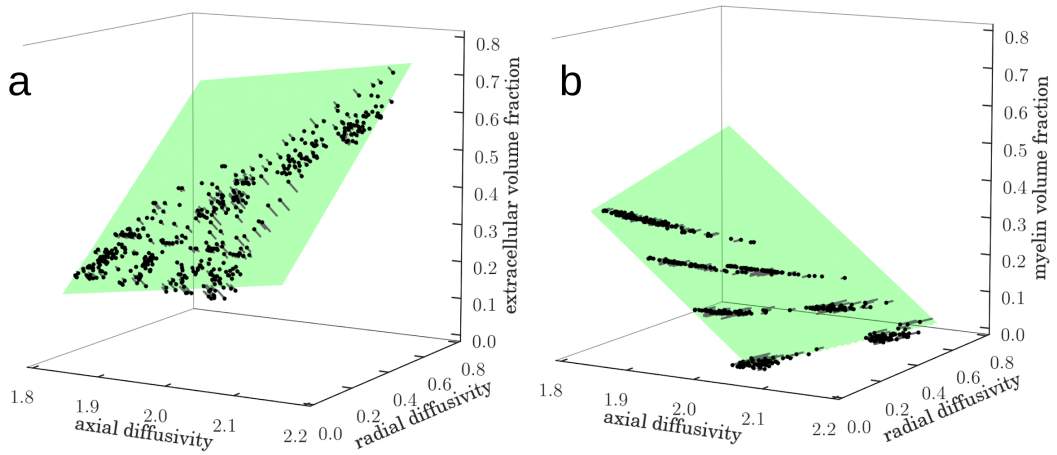


Figure 7.6.: Extracellular and myelin volume fractions as function of axial and radial diffusivities in units of 10^{-3}mm^2/s. Green surfaces minimise the orthogonal distances between data and the planes. Distances are shown as lines.

7.3.7. Dispersion

The simulated dispersion, which is effectively angular blurring of the signal (without interpolation of the signal), manifests as a reduction of axial and an increase in radial diffusivity, causing an overall reduction of FA. Figure 7.7 shows the effect of dispersion on FA, radial and axial diffusivity to be largely independent of packing density and demyelination simulation scenario.

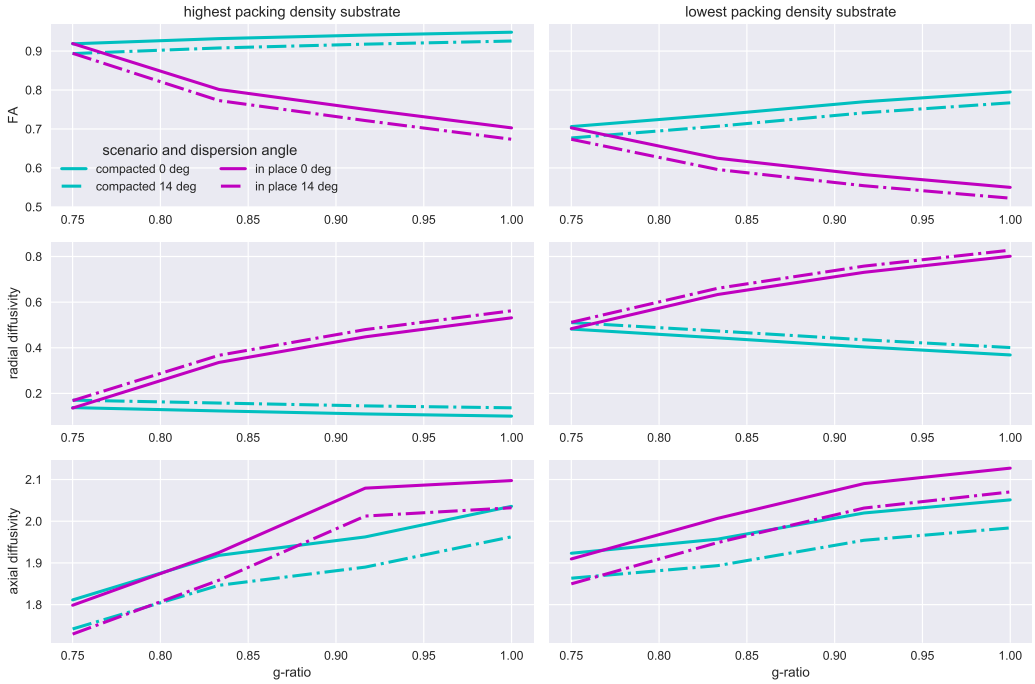


Figure 7.7.: Diffusion tensor measures for the substrates with the highest and lowest initial packing density undergoing demyelination in either scenario, with and without dispersion. Diffusivities in units of 10^{-3}mm^2/s.

7.4. Discussion

The demyelination Monte Carlo simulations investigate the influence of changing volume fractions during changes in myelin volume fraction on the predictive value of diffusion tensor indices for clinically relevant diffusion times and b-values. The effects of axonal packing density and myelin content on diffusion tensor quantities are often reported in clinical and preclinical studies; radial diffusivity in particular is often associated with myelin content or used as a marker for de- and remyelination.

In the developing brain with a given initial configuration of unmyelinated axons, an increase in myelin content could be accompanied by a decrease in extracellular volume fraction with axons remaining "in place"; or myelinating white matter tracts could expand

in cross-sectional area with constant extracellular volume fraction, at the expense of surrounding tissue or by pruning of existing axons (the inverse "shrinking" scenario).

This chapter investigates the specificity of the diffusion tensor indices on (de)myelination, taking into account what happens to the space that would have been taken up by myelin. Myelin could be replaced either by extracellular space, or alternatively the white matter tracts could compact down to take up this extra space (compare fig. 7.4). The former case may happen in the early stages of demyelination for instance, whereas the latter might be expected in the more chronic stages, or in unmyelinated tracts. Similarly, during white matter development, axonal pruning and changes to the extra-axonal space can coincide with myelination, dwarfing the effect of changing myelin content on the diffusion tensor indices and fundamentally rendering them non-specific to myelin. For example, one could ask whether findings based on DTI measures translate between transgenic myelin-deficient ("shiverer") mouse models [Readhead, Hood, 1990] and demyelinating diseases – even for perfectly parallel fibres.

7.4.1. Comparison with literature values

The simulation results for FA cover the range of reported in-vivo values for white matter of 0.87 for b=1000s/mm^2 and 0.92 for b=3000s/mm^2 [Tournier, Calamante, Connelly, 2013; Yoshiura et al., 2001] and yield diffusivities on the order of 0.78×10^{-3}mm^2/s, 1.9×10^{-3}mm^2/s, 0.23×10^{-3}mm^2/s for mean, axial and radial diffusivity, respectively. Reported values for mean, axial and radial diffusivity measured in the optic radiation at b=1000s/mm^2 are 0.82×10^{-3}mm^2/s, 1.3×10^{-3}mm^2/s and 0.57×10^{-3}mm^2/s, respectively [Klistorner et al., 2015]. For major white matter tracts implicated in language processing [Ivanova et al., 2016] those values are 0.83×10^{-3}mm^2/s, 1.2×10^{-3}mm^2/s and 0.64×10^{-3}mm^2/s, respectively. Axial and radial diffusivity measured at b = 700s/mm^2 in the posterior corpus callosum, are 1.58 ± 0.14^{-3}mm^2/s and 0.41 ± 0.05^{-3}mm^2/s, respectively [Kumar et al., 2013]. Literature values for mean diffusivity in white matter range from 0.62 to 0.79×10^{-3}mm^2/s [Helenius et al., 2002].

It is not possible to directly compare axial and radial diffusivities between the above simulations, which use perfectly parallel axons, and in-vivo tissue measurements with an unknown degree of dispersion, and the likely presence of other cell types (e.g. oligodendrocytes, astrocytes). One can expect that the lack of microscopic dispersion in the model causes the axial and radial diffusivity to be further from the mean diffusivity compared to in-vivo measurements, which is indeed the case. Simulated macroscopic intravoxel dispersion with a standard deviation of 14 degrees reduces this effect (fig. 7.7) but is not sufficient to explain the observed difference between in-vivo and simulated anisotropy.

Axial diffusivity increases with complete demyelination by approximately 17% for b=1000s/mm^2 and by 22% for b=3000s/mm^2. This increase can be attributed to the reduced volume fraction of myelin, which has lower diffusivity compared to the other compartments. This finding is in apparent contradiction with MS lesion studies that show no or only weak correlation with increased mean diffusivity [Fox et al., 2011; Mottershead et al., 2003]. Furthermore, axial diffusivity in shiverer mice is not correlated

with myelin content [Song et al., 2002]. This suggests that these changes could be masked in-vivo for instance by the presence of dispersion. Alternatively, the role of the myelin signal plays in our simulations might be exaggerated. Indeed, when the myelin signal is set to zero, mean diffusivity is independent of the myelin water fraction.

We can conclude that modelling (de)myelination purely as changing volume fractions is insufficient to adequately account for all aspects of in-vivo demyelination and associated encephalopathy related effects observed in in-vivo diffusion tensor imaging. In pathology such as oedema or inflammation, microstructural changes are expected to alter the diffusion properties of the tissue in ways that were not modelled here. They are a likely driving factor for the signal changes observed *in-vivo*.

7.4.2. Limitations

In other work, segmented electron micrographs were used to increase the realism of white matter simulations [Xu et al., 2015; Panagiotaki, Hall, Zhang, 2010]. Yet, recent Monte Carlo simulations suggest that round cylinders are a valid approximation at least for in-plane diffusion [Kleinnijenhuis et al., 2016]. Furthermore, the geometrical simplicity of my tissue model is a deliberate choice to account for the limited specificity of diffusion tensor indices in tissue with multiple fibre configurations.

For healthy mature tissue in the superior longitudinal fascicle of macaques, g-ratios range from 0.52 to 0.86 with a median of 0.74 [Liewald et al., 2014]. The theoretical maximal conduction velocity for neurons is achieved with a g-ratio close to 0.6 but it is generally higher in the central nervous system [Goldman, Albus, 1968]. Also, g-ratios vary depending on species and axon size [Hildebrand, Hahn, 1978; Waxman, Bennett, 1972]. To simplify the axon placement with gamma-distributed inner radii, I chose to ignore the axon diameter dependent g-ratio distribution as well as any dependence of the degree of demyelination on axon radius, as for instance reported for MS. The influence of non-axonal cells such as glial cells and oligodendrocytes on the diffusion signal is not well studied and not part of the simulations. These simplifications might have an impact on the outcome of the simulations and therefore the conclusions drawn.

The simulations are focused on changes in tissue geometry and disregard interaction of tissue compartments such as induced local gradients due to spatially and orientationally varying susceptibility. Spatially heterogeneous susceptibility due to incorporated gas can lead to underestimated diffusion coefficients [Hong, Thomas Dixon, 1992; Lian, Williams, Lowe, 1994] and, in white matter, the frequency of tissue boundaries varies heavily between axial and radial directions, which may influence measured anisotropy. However, Clark, Barker, Tofts showed that susceptibility-induced gradients do not significantly influence diffusion anisotropy at 1.5T [Clark, Barker, Tofts, 1999].

The simulated protons do not experience surface relaxation at axon or myelin boundaries and all simulated cell membranes are assumed to be impermeable. Our simulations do not take the effect of magnetisation transfer and change of T_2 due to exchange into account. This is not expected to be a major source of error as shown in bovine optic nerve [Stanisz et al., 1999]. Non-zero cell membrane permeability would increase the average mobility of water molecules in the radial direction, which would lead to an increase

in measured radial diffusivity. However, for typical intra-axonal pre-exchange times of 550ms [Quirk et al., 2003] for axons with a diameter of 2 μm, δ=30ms and Δ=30ms, this was shown to have only a minor impact on the measured radial signal [Raffelt et al., 2012].

7.5. Conclusion

Radial diffusivity is shown to be only indirectly dependent on the myelin content and is heavily confounded by changes in axon packing density during g-ratio changes. Changes in myelination can have divergent effects on radial diffusivity depending on the geometrical manifestation of the process. Changes in mean diffusivity and FA are mainly driven by changes in radial diffusivity. In the simulations, axial diffusivity increases with demyelination irrespective of simulation scenario and the degree of myelination can be determined using radial and axial diffusivity as linear regressors.

Using a simple geometric model of white matter undergoing changes of myelination, we show that radial diffusivity is not specific to myelin content but axial diffusivity might be. This is in agreement with recent findings using tissue-clearing and DTI [Chang et al., 2017]. In our simulations, myelin volume fractions can be predicted by jointly regressing radial and axial diffusivity, which is in apparent contradiction with in-vivo demyelination studies, indicating that demyelination may affect factors not included in the present simulations.

The simulations show that diffusion tensor indices are sensitive to myelin content but the changes in intra- and extracellular volume fraction during demyelination might outweigh the effect of demyelination on radial diffusivity and fractional anisotropy. The two simulated scenarios disentangle the effect of changes in extracellular ("in-place") or intra-axonal ("shrinking"') volume fractions during demyelination.

Recent *in-vivo* diffusion studies use the dependency of the diffusion coefficient on sequence parameters [Lee, Fieremans, Novikov, 2016] and fit multi-tensor and multi-tissue compartment models [Jelescu et al., 2016; Wang et al., 2015] to measure myelin volume fraction or use multiple modalities to estimate tissue volume fractions [Stikov et al., 2015a]. The current state of the field in *in-vivo* g-ratio acquisition and analysis techniques is reviewed in [Campbell et al., 2017]. Myelin water fraction-based g-ratio maps of babies born permaturely were presented in [Melbourne et al., 2014] and the first study of g-ratio measurements of children was reported in [Dean III et al., 2016]. Application and validation of the specificity of these techniques to myelination during brain development are promising but currently underexplored research endeavours.

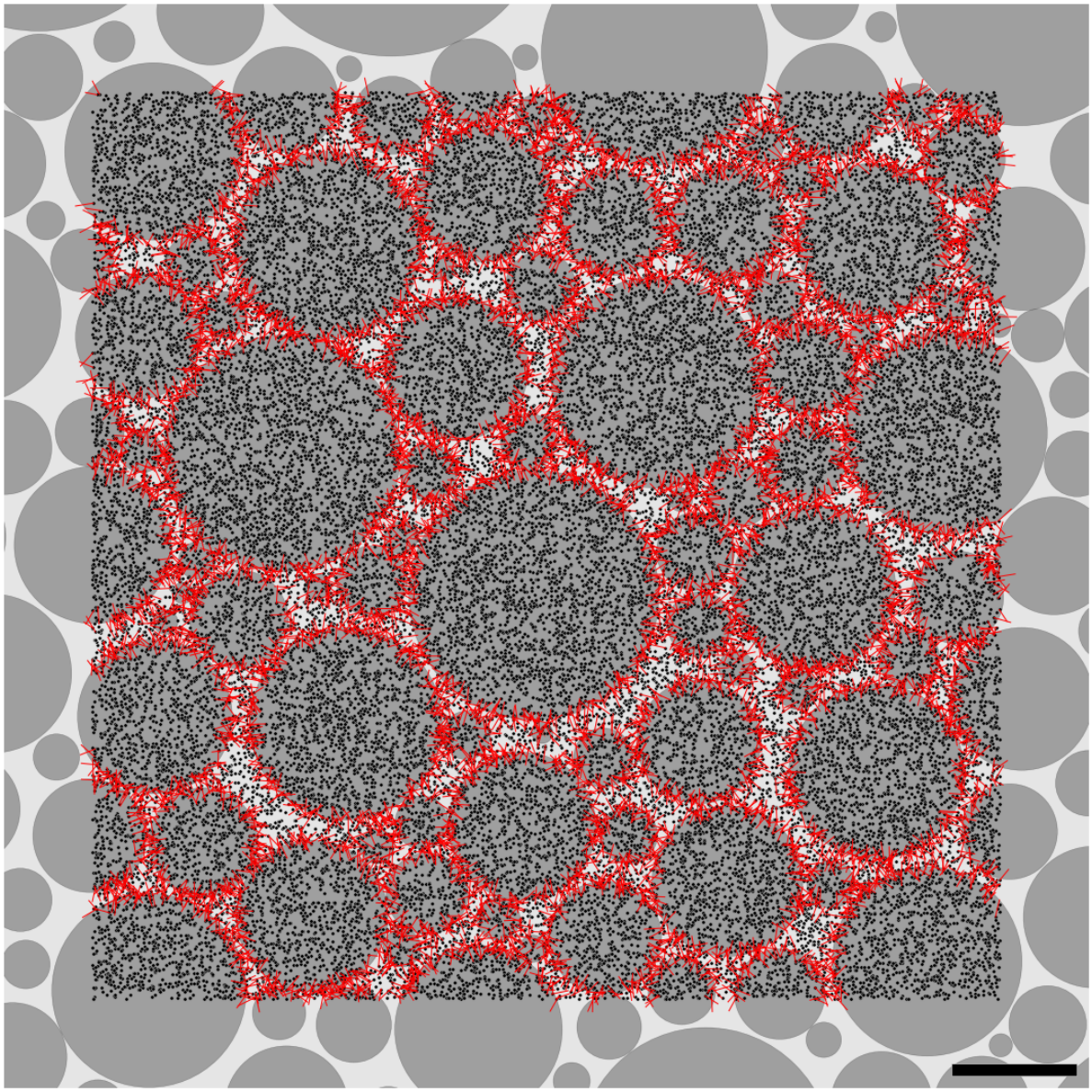


Figure 7.8.: Simulation of the worst-case scenario of the highest density substrate at the highest tissue compaction (no myelin): 19.6% of walkers experience a boundary in each time step. The image shows a magnified view of the substrate. Scale bar: 1μm, trajectories of walkers that interact with boundaries depicted in red.

Chapter 8

Multi-component neonatal brain HARDI template

Contents

8.1. Introduction

There is increasing interest in studying the developing brain using advanced multi-shell diffusion analysis methods, due to their potential to demonstrate microstructural features

not visible using other modalities. As reviewed in chapter 2, during the neonatal period, the human brain increases in size rapidly [Brody et al., 1987] and cerebral tissue undergoes rapid changes in cellular composition, density and water content [Dobbing, Sands, 1973]. These dramatic changes are reflected in MRI contrast and, in recent years, a number of publications have focussed on mapping anatomical or functional properties of the developing brain [Shi et al., 2014; Akazawa et al., 2016; Oishi et al., 2011; Avants et al., 2015; Dittrich et al., 2014; Schuh et al., 2014; Habas et al., 2009; Kuklisova-Murgasova et al., 2011; Serag et al., 2012a; Serag et al., 2012b].

Developments in diffusion MRI acquisition strategies now allow the routine acquisition of large amounts of data to be collected in relatively short periods of time [Larkman et al., 2001]. This makes it possible to acquire eloquent multi-shell High Angular Resolution Diffusion Imaging (HARDI) data in neonates within acceptable scan times, providing microstructural information about the developing white matter (WM) not available using other imaging modalities. To investigate this development, analyses require group-wise non-linear registration of multi-shell HARDI data over large numbers of subjects to a common group-average space (the template space).

Numerous diffusion-based templates have been created for adult populations featuring Diffusion Tensor Imaging (DTI) measures [Peng et al., 2009; Goodlett et al., 2009; Mori et al., 2008; Yushkevich et al., 2008] and HARDI data [Bouix, Rathi, Sabuncu, 2010; Patel et al., 2010; Yeh, Tseng, 2011; Varentsova, Zhang, Arfanakis, 2014].

Here, I describe a method for generating a high-quality multi-shell HARDI group template of the developing brain at term equivalent age, which forms the foundation for group and longitudinal analysis of brain development in normal and pathological cohorts. The creation of a multi-component template required the extension of existing registration techniques to integrate the population-specific tissue contrasts in the registration architecture. The template aligns orientation-resolved microstructural features of the population. In particular, the template resolves the age-specific anisotropy in grey matter and provides a contrast between signal that is similar to brain tissue and to free water that is not accessible with structural MRI.

The template forms the basis for longitudinal modelling described in chapter 9. The multi-contrast registration of ODFs and the template were presented in [Pietsch et al., 2017a; Pietsch et al., 2017b].

8.2. Background

8.2.1. Image registration

The goal of medical image registration is to align and optionally deform ("warp") an image to another image, which enables analysis of corresponding features in a common reference space or analysis of the mapping to extracted morphometric information. Discussing the field of registration methods, the various cost functions employed, and their implicit and explicit constraints and regularisation is beyond the scope of this work. Here, the focus is on methods used for the creation of the multi-contrast template. For reviews of medical image registration in general see [Oliveira, Tavares, 2014; Viergever et al., 2016], and

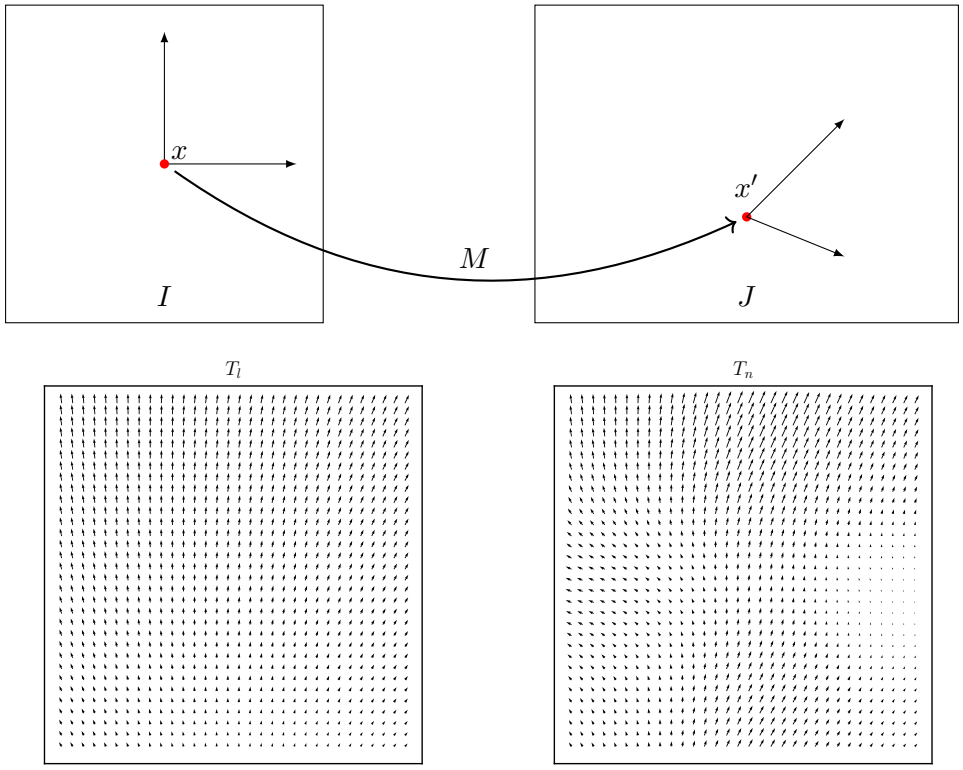


Figure 8.1.: 2D illustration of the coordinate transformation M from image I to image J (top) and a linear (T_l) and non-linear (T_n) displacement field defined in the space of I. A non-linear transformation in general does not preserve angles and distances.

for reviews focused on non-linear registration and recent developments such as slice to volume registration see [Crum, Hartkens, Hill, 2004; Andersson, Jenkinson, Smith, 2007; Sotiras, Davatzikos, Paragios, 2013; Ferrante, Paragios, 2017].

Finding an optimal mapping between two images (denoted M) involves minimising a cost function $\mathcal{C}$ that quantifies the difference between an image I and a second image J, subject to regularisation $\mathcal{R}$ that encourages plausible deformations

$$E = \mathcal{C}\left(I, J \circ M\right) + \mathcal{R}(M) \tag{8.1}$$

$J \circ M$ denotes the composition of the functions J and M, the former being a function that maps coordinates to intensity values (an image). The transformation M maps every coordinate $\mathbf{x}$ in the reference space (I) to the corresponding location $\mathbf{x}'$ in J (see fig. 8.1).

$$M : \mathbf{x} \rightarrow M(\mathbf{x}) = \mathbf{x}' \tag{8.2}$$

Note that evaluating registration accuracy is challenging due to multiple sources of errors that affect E [Zitova, Flusser, 2003] and the competing effects of minimising the dif-

ference between images and obtaining plausible and useful deformations via the choice of $\mathcal{R}$ and the distance measure $\mathcal{C}$. However, it is an ill-posed problem to use the transformed images alone to assess the quality of the alignment [Rohlfing, 2012] as it always depends on the subsequent analysis. Without regularisation it is possible to map images so that they resemble each other to arbitrary degree, which would remove any intensity-based differences between subjects and make them useless for image intensity-based analysis, as demonstrated in the "Completely Useless Registration Tool" [Rohlfing, 2012]. A 'worse' spatial alignment could preserve more of the intensity differences between both images. Similarly, in pathology such as tumours or lesions, desirable criteria for alignment might not be well defined or different from those for healthy tissue. The choice of algorithm and metric can differ in the presence of severe pathology or large developmental or age differences between images and can impact the specificity of downstream analysis.

Although there are many distance measures available, for intensity-based registration, a reasonable distance measure is the mean squared intensity differences calculated across the area of overlap Ω

$$\mathcal{C}_2(I, J, \circ M) = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} (I(\mathbf{x}) - J(M(\mathbf{x})))^2 \tag{8.3}$$

The best transformation could be found by grid search if the transformation has few parameters but in practice it is the result of a form of gradient-based optimisation that minimises the cost function eq. (8.1) using the gradient with respect to the transformation parameters $\partial E(I, J, M)/\partial M$. As in gradient-based learning (discussed in section 4.2.2) gradient descent algorithms can be used to optimise the transformation via minimisation of E.

8.2.1.1. Transformation representations

Registration algorithms can be grouped by the type of transformation used. Rigid transformations are limited to global translation and rotation and therefore preserve angles and distances. Affine transformations allow an additional global shear and scaling of the image. Affine and rigid transformations are referred to as linear transformations, while non-linear transformations typically have a much higher degree of freedom to express spatially varying transformations (deformations).

Linear transformations can be applied compactly to any coordinate $\mathbf{x}$ via a dot product of an affine 3x3 matrix ($\underline{A}$) with $\mathbf{x}$ and an addition of the translation vector $\mathbf{t}$

$$\mathbf{x}' = \underline{A}\mathbf{x} + \mathbf{t} \tag{8.4}$$

When $\mathbf{x}$ is parametrised in homogeneous coordinates, the linear transformation is

summarised in a single 4x4 transformation matrix $\underline{L}$

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} A_{11} & A_{12} & A_{13} & t_1 \\ A_{21} & A_{22} & A_{23} & t_2 \\ A_{31} & A_{32} & A_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{\underline{L}} \begin{pmatrix} x_2 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} \tag{8.5}$$

Non-linear transformations are either represented using sparse control point-based deformations (such as free-form deformations [Rueckert et al., 2006]) or deformations sampled densely on the voxel-grid, also referred to as warps. A displacement is the distance vector between the feature location in I and the corresponding feature in J and is defined on the image grid of M: $T(\mathbf{x}) = \mathbf{x}' - \mathbf{x}$. See fig. 8.1 for two example displacement fields.

A mapping should be topology-preserving; if it is possible to map each location in I to a location in J, then it is reasonable to expect that each location $\mathbf{x}'$ originates from a unique location $\mathbf{x}$. Furthermore, a transformation that maps the anatomical or functional representation of one subject to that of another subject is expected to be relatively smooth compared to the voxel grid. A mapping between two manifolds that is differentiable (smooth) and has a differentiable inverse (one-to-one mapping) is referred to as a diffeomorphism [Arnold, Khesin, 1992]. A diffeomorphic mapping can be generated by step-wise composition of small diffeomorphic warps [Cootes et al., 2004]. Note that the assumption of a diffeomorphic mapping is violated for instance when mapping non-matching cortical folding patterns or normal to abnormal brains. However, in practise, diffeomorphic algorithms outperform other free form registration algorithms on human brain data [Klein et al., 2009]. See [Holden, 2008; Sotiras, Davatzikos, Paragios, 2013] for an overview of diffeomorphic algorithms.

8.2.2. Symmetric diffeomorphic registration of ODFs

The demons algorithm [Thirion, 1998] uses a diffusion model to iteratively apply local forces to the mapping, gradually and smoothly changing the transformation. In the formulation above, this algorithm can be expressed as a step-wise minimisation of the loss function with a regularisation on the *change* in the mapping, followed by Gaussian blurring of the deformation field [Vercauteren et al., 2009; Hernandez, Olmos, Pennec, 2008]. Vercauteren et al. extended the demons algorithm to use an update rule that ensures a diffeomorphic mapping. However, the loss function used, similar to eq. (8.1), was not symmetric with respect to I and J. The intensity value of $J(\mathbf{x}')$ is obtained through interpolation but that of $I(\mathbf{x})$ without interpolation. Asymmetric smoothing yields biased results of either under- or overestimation of change, depending on which image is kept fixed [Yanovsky et al., 2008] and swapping the images will not yield transformations that are inverse-consistent. A lack of inverse consistency between the mapping from I to J (M_{IJ}) and the reverse transformation (M_{JI}) can lead to tissue compaction in one image that is not matched by tissue expansion in the other image.

This can be prevented by using a symmetrically parametrised transformation that prohibits or penalises inverse-inconsistency [Avants et al., 2008]. For unbiased interpolation,

the distance measure can be evaluated on a grid in the space that lies halfway between I and J (Ω_h)

$$\mathcal{C}_{2,\text{sym}}(I \circ M^{-\frac{1}{2}}, J \circ M^{\frac{1}{2}}) = \frac{1}{|\Omega_h|} \sum_{\mathbf{x} \in \Omega_h} \left(I(M^{-\frac{1}{2}}(\mathbf{x})) - J(M^{\frac{1}{2}}(\mathbf{x})) \right)^2 \tag{8.6}$$

Diffusion images carry orientationally resolved information about the tissue microstructure. Therefore, non-linear warps applied to diffusion weighted MRI (dMRI) volumes need to take the proper reorientation, scaling and shear of the tissue into account. For instance, applying a local shear to an isotropic diffusion profile would make it anisotropic and fundamentally change the interpretation of the microstructure. Early work focused on the transformation of the eigenvectors of diffusion tensors [Zhang et al., 2006; Yeo et al., 2009]. Unfortunately, diffusion tensor models cannot resolve crossing fibres nor represent the rich microstructural information in HARDI (see section 3.4.2). To address this, registration methods were developed to align diffusion attenuation profiles represented in spherical harmonics (see section 3.5.3) [Geng et al., 2011; Bloy, Verma, 2010] or as Gaussian mixture models [Cheng et al., 2009]. Yet, a model-free representation of the data can not appropriately preserve the continuity of directions and cross-sectional area of white matter tracts across voxels [Zhan, Yang, 2006; Tournier et al., 2008; Raffelt et al., 2011; Raffelt et al., 2012].

Raffelt et al. proposed an extension of the symmetric diffeomorphic demons algorithm of [Avants et al., 2008] to white matter fibre orientation distribution functions represented in spherical harmonics, taking the appropriate reorientation of ODFs into account. This method provides an improved alignment over diffusion tensor-based registration for subsequent group analysis of white matter pathologies [Raffelt et al., 2011] and can be used to investigate morphological changes specific to white matter bundles [Raffelt et al., 2012].

The metric driving the registration is the mean squared difference in the spherical harmonics coefficients between both images after reorientation. When I and J are expressed in the basis of real spherical harmonics of degree l and order m, the squared intensity difference on the right hand side of eq. (8.6) simply becomes the sum of the squared difference in the coefficients [Raffelt et al., 2011]

$$\left(I(M^{-\frac{1}{2}}(\mathbf{x})) - J(M^{\frac{1}{2}}(\mathbf{x})) \right)^2 = \sum_{l=0,2,\ldots}^{l_{\max}} \sum_{m=-l}^{l} \left(I(M^{-\frac{1}{2}}(\mathbf{x}), l, m) - J(M^{\frac{1}{2}}(\mathbf{x}), l, m) \right)^2 \tag{8.7}$$

As described in detail in [Avants et al., 2008], registration is performed iteratively via gradient descent and regularisation is applied by smoothing the cost function gradient field and the total displacement field. Registration proceeds in an iterative process until convergence is achieved in steps with increasing spatial resolution and increasing angular frequency terms.

8.2.3. Unbiased cross-sectional template creation

The purpose of creating a population-specific diffusion template is to define a single representative map of anatomical and microstructural features. A single template might

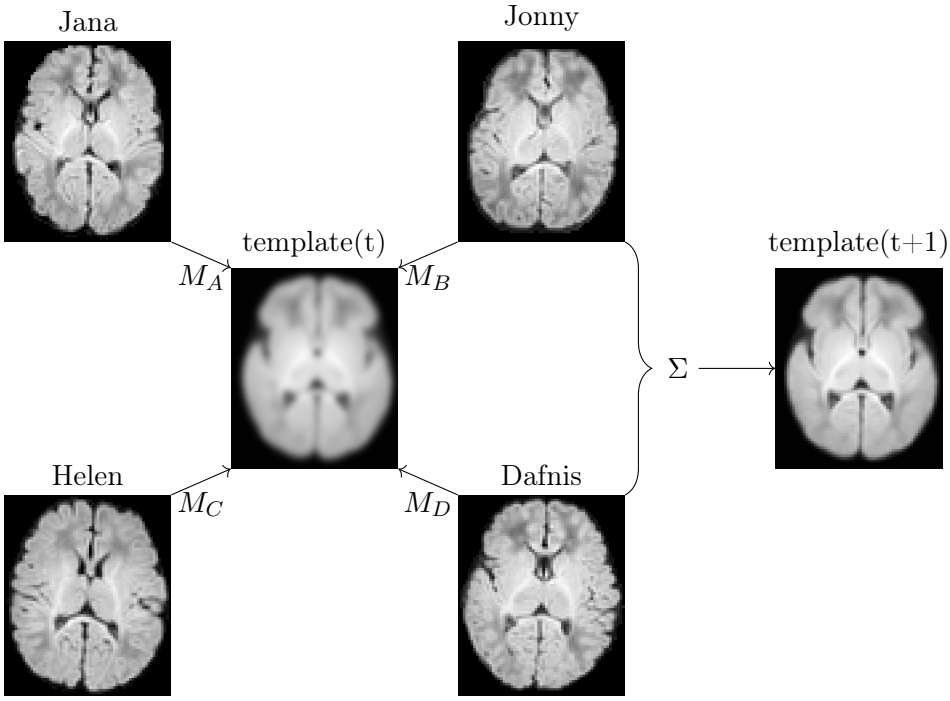


Figure 8.2.: Illustration of an iterative intensity-based template creation method. Using an initial template, each subject is registered and aligned to the template space. The transformed images are averaged and become the template for the next iteration $(t + 1)$.

not be ideal for applications, such as segmentation [Iglesias, Sabuncu, 2015] or information propagation between subjects [Cardoso et al., 2015] but aligning all subjects to a common space facilitates volumetric analysis and comparison of microstructural features between groups. Aligning and averaging images removes subject- and image-specific variability or noise and allows investigating group-differences on a per-voxel and, in the case of HARDI data, per-fibre basis. The resulting subject-to-template warps can be used for further analysis of morphology and microstructural tissue properties, such as tract-specific differences in cross-sectional area [Raffelt et al., 2011].

There are many heuristics for creating an average representation of a population [Guimond, Meunier, Thirion, 2000; De Craene et al., 2004; Joshi et al., 2004; Bhatia et al., 2004; Lorenzen, Davis, Joshi, 2005; Park et al., 2005; Zöllei et al., 2005; Commowick, Malandain, 2006; Avants et al., 2010; Reuter et al., 2012]. Approaches differ in the choice of stereotactic space (single subject, reference space, population average), in the type of transformation that maps subjects to that space, and in the way the average shape and appearance is calculated. See [Evans et al., 2012] for a review about brain templates and techniques.

A template can be constructed by a combined optimisation of all subjects' shape and intensity to a common space [Studholme, Cardenas, 2004] but this approach is

computationally very demanding. Typically, population templates are created by either averaging the appearance (intensity) [Joshi et al., 2004; Lorenzen et al., 2006], or the shape (deformations) [Vaillant et al., 2004; Younes, 2007; Beg, Khan, 2006].

In [Avants et al., 2010], a hybrid approach is proposed that first estimates a mapping between subjects and an initial template. This is followed by an optimisation of the average appearance by minimising the total pairwise distance in shape-space between subjects and the initial template. Finally, the template appearance is optimised. Avants et al. show that the explicit optimisation of shape has a small but significant effect on the accuracy of aligning the hippocampus of diseased populations therefore increasing detection power of intensity and shape difference. However, in healthy populations both methods are practically equivalent.

Raffelt et al. create a population average based on registration to the average intensity image. They initialise the transformations between subjects and the average space using affine registration of Fractional Anisotropy (FA) maps to a common space using a block-matching registration approach [Ourselin et al., 2001]. The WM orientation distribution functions (ODFs) are transformed to that space and averaged to build an initial template (see fig. 8.2). This initial template is iteratively updated by registering each subject to the current template, transforming all images using the updated warps, and averaging of the aligned images. The template becomes sharper in the next iteration if a part of any image gets better aligned with the template. This is repeated until the template convergences.

8.3. Multi-contrast ODF registration for template creation

8.3.1. Extension to multi-contrast ODF registration

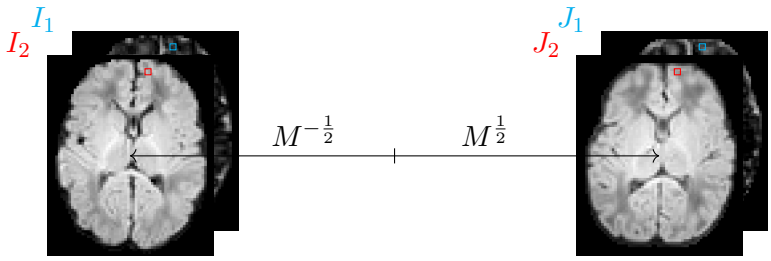


Figure 8.3.: Simultaneous symmetric registration of multiple contrasts (red and cyan) of images I and J. The cost is evaluated within contrasts but both contrasts drive the registration simultaneously.

Registration driven by WM fibre ODFs is powerful as it also provides, besides a spatially resolved white matter density map, information about the orientation of the tissue. However, areas with little or no white matter do not contribute to the alignment of the tissue. Even if a downstream analysis is focused on WM only, grey matter (GM) or cerebrospinal fluid (CSF) components are perfectly aligned with the WM component and can provide valuable information for inter-subject registration. Using only part of the tissue

contrast or ignoring parts of the brain for alignment might bias the registration. This is particularly relevant in neonates, where overlap between components is much more pronounced than in adults. Therefore, in [Pietsch et al., 2017a], we extended the ODF registration framework of [Raffelt et al., 2011] to simultaneously align multiple image contrasts.

This can be achieved by adapting the distance function eq. (8.6) to a weighted sum of contributions from component-specific image pairs I_c and J_c (see fig. 8.3)

$$\mathcal{C}_{2,\mathrm{sym},c}(I \circ M^{-\frac{1}{2}}, J \circ M^{\frac{1}{2}}) = \frac{1}{|\Omega_h| \sum_{c'} w_{c'}} \sum_{\mathbf{x} \in \Omega_h} \sum_{c} w_c \left(I_c(M^{-\frac{1}{2}}(\mathbf{x})) - J_c(M^{\frac{1}{2}}(\mathbf{x})) \right)^2 \tag{8.8}$$

By incorporating a component-specific weight w_c in the cost function, the cost function gradient become explicitly weighted by component. The parameter update in linear registration is proportional to the cost function gradient, which explicitly makes the linear transformation update weighted by component. Similarly for non-linear registration, the displacement field update is proportional to the gradient, hence the resulting total displacement field is contrast-weighted. Note that the transformation M is applied and optimised simultaneously for all components. In case of one or multiple orientation distribution function (ODF) components, reorientation is performed for each component separately.

8.3.2. Extension of the linear registration for the template creation

In contrast to [Raffelt et al., 2011], the initial linear registration is performed not on FA images but on ODF images, using a symmetric least-squares metric. Instead of block-matching, registration proceeds in multi-resolution stages starting with rigid followed by affine registration. Hence, the linear template is created analogously to the iterative non-linear registration (see fig. 8.2 and table 8.1). However, the average rigid or affine transformation is factored out from each subject-to-template transformation to ensure that the template remains centred between iterations. By forcing the average rigid or affine transformation to be the identity transformation, the population template remains representative in size and shear and does not drift or rotate between template iterations.

The matrix average $\underline{L}_{av}$ is calculated in the log-domain (log-Euclidean mean) using matrix logarithms and exponentials [Arsigny et al., 2009; Cheng et al., 2001]

$$\underline{L}_{av} = \exp\left(\sum_{i}^{N} \frac{1}{N} \log(\underline{L}_i) \right) \tag{8.9}$$

Each transformation matrix L_i is left-multiplied by the inverse average transformation $\underline{L}_i \leftarrow \underline{L}_{av}^{-1} \underline{L}_i$. Note that the log-Euclidean mean does not guarantee that the average matrix is rigid even if all matrices $\underline{L}_i$ are rigid. Therefore, for rigid registration, $\underline{L}_{av}$ is decomposed into the product of a scaling and a rotation matrix and the scaling matrix is factored out of $\underline{L}_{av}$ to only correct for rotation and translation without introducing scaling and shearing into the rigid template.

stage	type	scaling	steps	$l_{\max}$
0	align centres of mass	1	1	0
1	rigid	0.3	max 100	2
2	rigid	0.4	max 100	2
3	rigid	0.6	max 100	2
4	rigid	0.8	max 100	4
5	rigid	1.0	max 100	4
6	rigid	1.0	max 100	4
7	affine	0.3	max 500	2
8	affine	0.4	max 500	2
9	affine	0.6	max 500	2
10	affine	0.8	max 500	4
11	affine	1.0	max 500	4
12	affine	1.0	max 500	4
1	non-linear	0.3	5	2
2	non-linear	0.4	5	2
3	non-linear	0.5	5	2
4	non-linear	0.6	5	2
5	non-linear	0.7	5	2
6	non-linear	0.8	5	2
7	non-linear	0.9	5	2
8	non-linear	1.0	5	2
9	non-linear	1.0	5	4
10	non-linear	1.0	5	4
11	non-linear	1.0	5	4
12	non-linear	1.0	5	4
13	non-linear	1.0	5	4
14	non-linear	1.0	5	4
15	non-linear	1.0	5	4
16	non-linear	1.0	5	4

Table 8.1.: Linear and non-linear iterations for the creation of a population template. The initial template is updated after each stage. 'Scaling' refers to the spatial down-sampling of the individual images with respect to the original image grid size; $l_{\max}$ denotes at which order the spherical harmonics are truncated to control the angular resolution of the ODFs. For each subject to template registration of the linear stages, the transformation is optimised using up to 100 or 500 gradient descent steps, non-linear transformation are always updated using 5 steps.

8.3.3. Pairwise registration accuracy experiment

To investigate the effect different microstructural tissue-contrasts have on the registration accuracy, we used 20 minimally preprocessed HARDI datasets from the Human Connectome Project (HCP) cohort [Glasser et al., 2013]. The diffusion data has a spatial resolution of 1.25mm isotropic and is sampled on 4 shells b=5, 1000, 2000, and 3000s/mm^2 in 90 directions per shell with a TE of 89ms and a TR of 5.5s. See [Uğurbil et al., 2013] for acquisition details.

The images were bias field corrected using ITK's *N4* algorithm [Tustison et al., 2010] and intensity normalised using MRtrix' *mtnormalise* [Raffelt et al., 2017]. Using a data driven method [Dhollander, Raffelt, Connelly, 2016], we estimated tissue type response functions for WM, GM and CSF for each subject. Using the cohort's average responses, all images were deconvolved into 3 tissue-specific components using multi-shell multi-tissue constrained spherical deconvolution (MSMT-CSD) [Jeurissen et al., 2014] (see section 3.5.4). See fig. 8.4 for an exemplary tissue density map and WM ODFs.

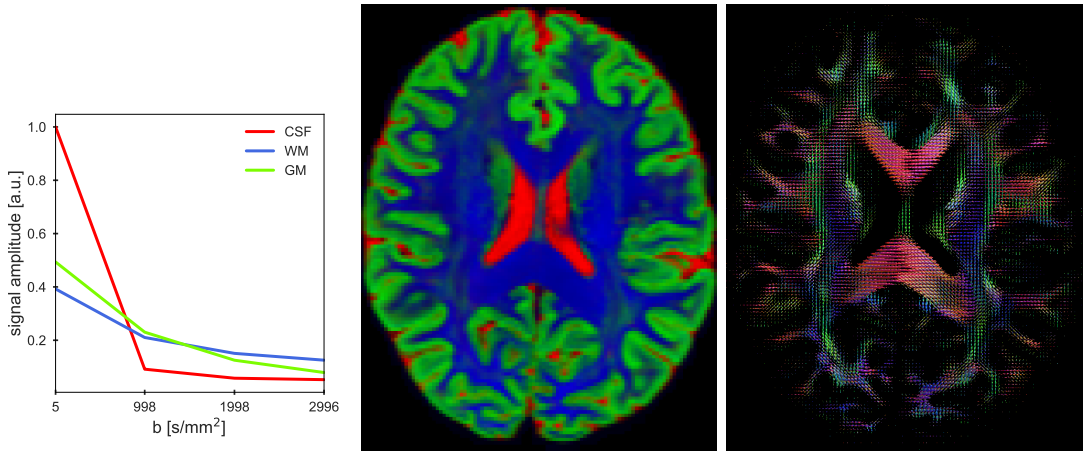


Figure 8.4.: Group average DC signal for each component (left), component density map (middle), and WM ODF image (right) of a single HCP dataset.

To assess the influence of the WM and GM components on the registration, we created a warped version of each set of images, and registered it back using various relative weights between the components, and assess the residuals between the original and back-transformed images. A realistic warp is created by registering the WM component of subject A (A_w) to that of a randomly chosen other subject B. The resulting transformation M_{AB} is used to warp the WM and GM components of A to the space of B:

$$A_w^B = A_w \circ M_{AB}$$
$$A_g^B = A_g \circ M_{AB}$$

Finally, the pair A_w^B and A_g^B is registered to A_w and A_g and transformed back to the

space of A using the resulting transformation $M_{A^B A}$:

$$A_w^{BA} = A_w^B \circ M_{A^B A}$$
$$A_g^{BA} = A_g^B \circ M_{A^B A}$$

If the last registration was perfect, it would undo the first transformation M_{AB}. Hence, to judge the quality of the registration, one could assess the distance of $M_{AB} \circ M_{A^B A}$ from the identity transformation. This, however does not give tissue resolved information about the quality of the registration. An alternative approach is to calculate the difference between the images A and A^{BA} within the brain mask Ω_m for a specific contrast using the absolute intensity difference

$$d_c(A, A^{BA}) = \frac{1}{|\Omega_m|} \sum_{\mathbf{x}} |A_c(\mathbf{x}) - A_c^{BA}(\mathbf{x})| \tag{8.10}$$

or using the root mean squared intensity difference for multiple contrasts

$$d(A, A^{BA}) = \frac{1}{|\Omega_m|} \sum_{\mathbf{x}} \sqrt{\frac{1}{\sum_{c'} 1} \sum_{c} \left(A_c(\mathbf{x}) - A_c^{BA}(\mathbf{x})\right)^2} \tag{8.11}$$

By changing the relative tissue weight of the WM component in the last registration ($M_{A^B A}$), it is possible to investigate whether GM can contribute to the registration accuracy and how to weight the components. Note that transforming and resampling image A to the space of B and back causes image blurring and therefore non-zero residuals. For comparison reasons, residuals are normalised by division with the lowest residual of any weighting.

Figure 8.5 shows the effect of the weightings of the WM component relative to the GM component on the voxel-wise residuals for the intra-subject registration experiment. Weights of both components are normalised to sum to 1, hence 0.5 represents equal weighting. Registration driven only by the WM component yields lower WM residuals compared to registration driven solely by the GM component. Across the brain, GM residuals are comparable for those two scenarios but deeper GM (eroded mask) benefits more from the GM component. Consequently, WM does not only align WM but also aids in aligning superior GM.

In general it is beneficial to include both components for the alignment of either component. Optimal weights differ, depending on which contrast's residuals need to be minimised and which region of the brain is considered and there is a clear trade-off between using either contrast. The weights could be tuned if the analysis was specific to one component or localised to specific areas of the brain, yet, across the brain, equal weighting between both components is close to optimal for the alignment of both WM and GM.

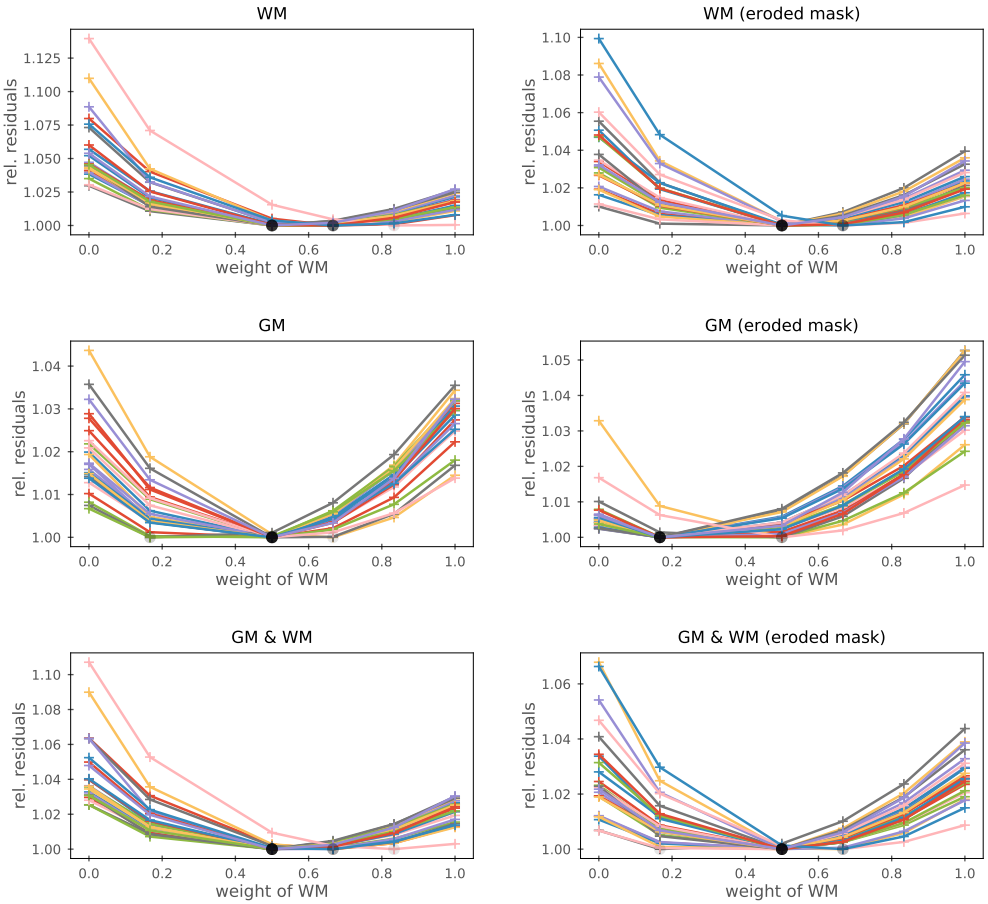


Figure 8.5.: Normalised average residuals after transformation of each HCP image onto another subject and subsequent registration with the original undistorted image. Plotted lines show the residual of the WM, GM and both components of each subject for varying relative weighting of the WM component to the registration. The plots in the second column display the residuals evaluated on a 10-times eroded brain mask excluding most of the cortex. Semi-transparent black circles indicate the lowest residual for each subject.

8.3.4. Group template experiment

The experiment of registering a distorted image back to its original state is unlikely a practical application of multi-contrast ODF registration. To investigate the effect of additional tissue contrast on a population template, we created two templates of the 20 HCP datasets using an identical procedure: rigid, followed by affine registration (section 8.3.2), and finally by non-linear registration (method of Raffelt et al., section 8.2.3) to iteratively create templates with increasing spatial and angular definition. The spatial and angular resolutions of each stage are outlined in table 8.1.

One template was created using only the WM components driving registration (denoted W). For the other template both WM and GM components were registered jointly with each contrast weighted equally (denoted C). To compare the templates visually on the same image grid, they were registered and transformed to their common midway space using the respective WM components (W_w and C_w) for registration.

Both templates' WM ODFs appear very similar in terms of spatial sharpness and tract orientations. Figure 8.6 shows the WM component of both templates in a region of the superior WM and cortex in sagittal projection close to the centre of the brain (see fig. 8.7 a for an overview image). The template using both contrasts (C) has a slightly higher WM density (warm colours) within white matter regions and a lower density outside these regions, indicating a better separation of cortical WM from GM.

Note that using the sharpness of a template as a surrogate for registration quality is reasonable to some extent but has limitations. The spatial arrangement of tissue microstructure, WM tracts and cortical folding patterns are remarkably consistent between subjects on a coarse scale but vary in shape, size, and location on a finer level [Ronan, Fletcher, 2015; White et al., 1997; Lohmann, Cramon, Steinmetz, 1999; Thompson et al., 1996]. A single template using diffeomorphic registration can not generate transformations that consistently align structures such as the Heschl's gyrus in the auditory cortex, which is duplicated in up to 60% of the population [Leonard et al., 1998; Evans et al., 2012].

In connectomics, it is desirable to map white matter tracts to cortical GM regions. Determining an accurate and unbiased termination point of WM fibres is challenging and tractography algorithms are biased to end in the gyral crown, where fibres follow nearly straight trajectories (up to 6 degree bending per 400 μm), compared to steeper angles at the sulcal walls (22 to 49 degree in 400 μm), where fibres turn by up to 90 degrees within less than 1.5mm [Schilling et al., 2018].

Blurring due to poor spatial separation of WM fibres at the sulcal wall causes an averaging of fibres oriented along the WM tract and those bending into the cortex. Figure 8.7 compares the principal direction of WM ODFs of both templates. In the combined template, WM ODFs bending into the cortical GM have a slightly higher curvature than in the template created using only the WM component (white quivers in fig. 8.7 d) and multi-contrast registration produces more orthogonal ODFs in adjacent voxels, consistent with findings in high-resolution post-mortem scans [Miller et al., 2011].

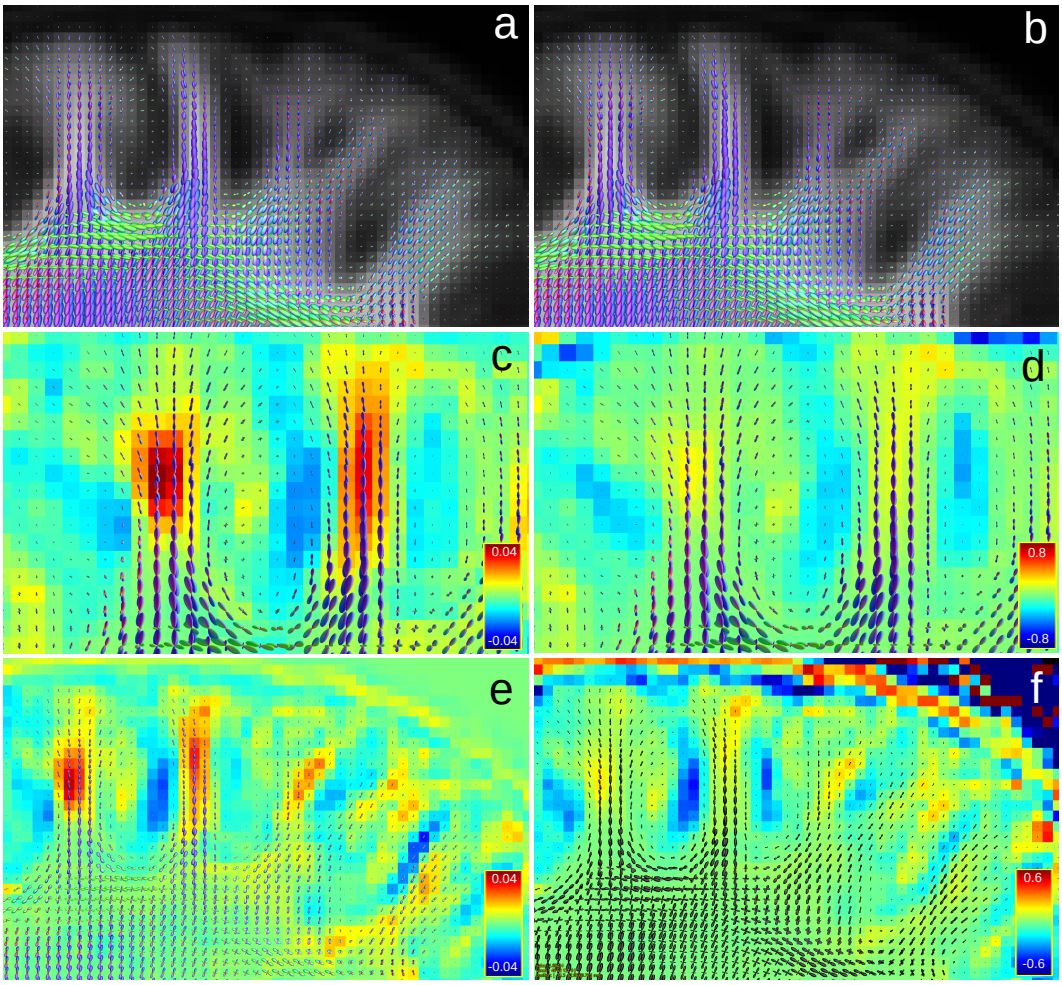


Figure 8.6.: Higher contrast between high and low density WM ODFs areas in the cortex of the HCP template generated using WM and GM (C) compared to the WM only driven registration of template W. (a) WM ODFs overlaid onto the WM density for the W_w template and (b) for the C template. Absolute (c,e) and relative (d,f) difference in WM density between both templates. Warm colours indicate higher density in the C template. In figures c - f, ODFs are of the C template.

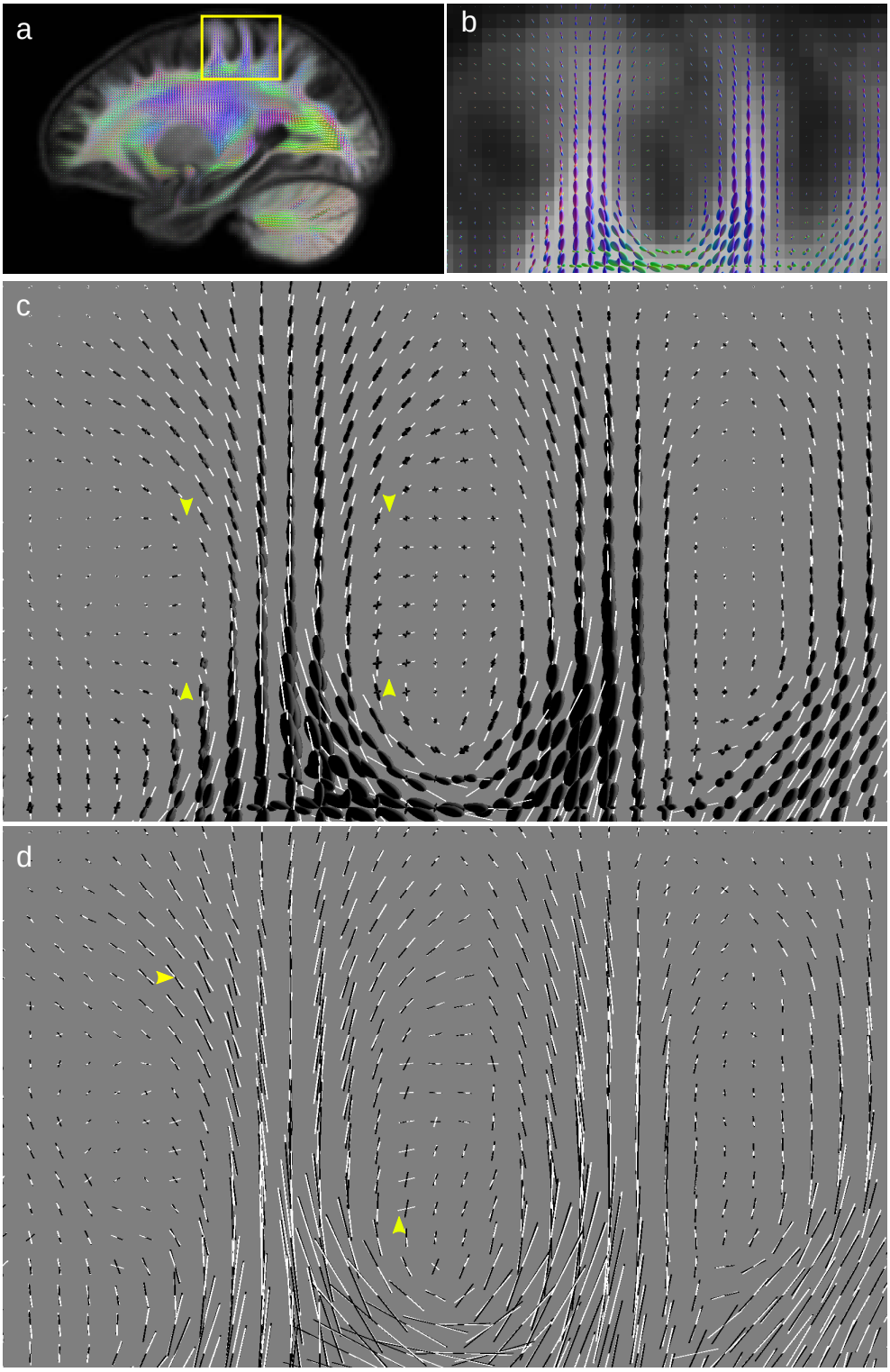


Figure 8.7.: Comparison of the direction of fibres projecting into the cortex of the template generated using only the WM contrast (W) and that of the combined WM and GM template (C). (a) Overview of the WM ODFs (W template). (b) Zoomed in view of the cortex with the WM density image shown as the background. (c) WM ODFs (black) of the C template with principal direction quivers scaled by WM density (white). (d) Principal directions of C (white) and principal directions of W (black). The arrows in (c) point at areas where ODF principal direction change by a steep angle. Principal directions of the the W template (black) tend to follow the main WM bundle and exhibit lower curvature (arrows in d).

8.3.5. Conclusion

HARDI data provides unique microstructural contrast that, using MSMT-CSD, can be separated into tissue components. In contrast to multi-modal registration, these components are perfectly aligned and they naturally complement each other in spatial and orientational information.

We have shown that including WM ODFs and scalar GM images simultaneously in the registration metric improves the accuracy of the registration and produces sharper delineations between WM and GM in the cortex. The inclusion of the GM tissue types seems to provide a moderate improvement in the alignment of subcortical white matter indicated by slightly increased curvature of fibres bending into the cortex (fig. 8.7). The optimal weights to assign to each tissue type need to be determined based on the target application and importance of WM alignment relative to GM alignment but an equal weighting is a sensible default.

8.4. Neonatal template

8.4.1. Introduction

In adults, WM, GM and CSF have distinct diffusion signal profiles across shells (see fig. 8.4), which is the main feature allowing a decomposition of the HARDI signal into tissue-specific components. In neonates, however, the mean signal decays similarly in cortical grey matter as in white matter structures such as the corpus callosum (CC) (see fig. 8.9). Furthermore, cortical grey matter voxels exhibit a high degree of radial organisation, remnants of migratory and ongoing developmental processes described in section 2.2.3. Hence a separation based on anisotropy is difficult at best. Also, mean signal decay curves are less coherent between the body and the genu of the CC than between WM and GM. Therefore, due to the ambiguities in differentiating between WM and GM in this age range, the focus for this study is on separating the anisotropic WM-like signal from the CSF-derived isotropic 'free water' component. Note that this decomposition uses signal characteristics found in WM and CSF to decompose the images. Due to overlapping signal characteristics, it does not provide a tissue separation in the biological sense but a separation into images that best fit the chosen 'free water' and 'tissue' component fingerprints.

8.4.2. Cohort and preprocessing

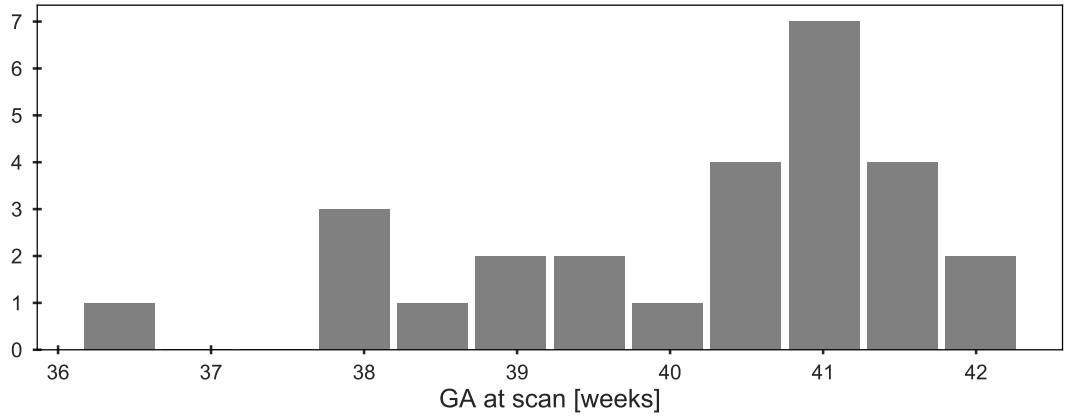


Figure 8.8.: Age distribution of the Developing Human Connectome Project (dHCP) cohort.

The cohort consists of 27 healthy term control babies acquired as part of the dHCP. The youngest subject has a postmenstrual age at scan of 36.1 weeks, the oldest 42.3 weeks; the average gestational age at scan is 40.2 weeks (see fig. 8.8).

The multi-shell high angular resolution diffusion single-shot spin-echo echo-planar images were acquired on a Philips 3T Achieva scanner using a dedicated neonatal head coil [Hughes et al., 2017a] with a maximum gradient amplitude of 70mT/m. The 300 volumes per data set were sampled with four phase-encode directions on four shells with b-values

of 0 (n=20), 400 (n=64), 1000 (n=88) and 2600 (n=128) with TE=90, TR=3800 ms [Tournier et al., 2015a; Hutter et al., 2017] and reconstructed to a resolution of 1.5 mm. See section 6.3.1 and [Hutter et al., 2017; Tournier et al., 2015b] for details about the acquisition and optimisation of contrast for neonatal imaging.

The images were preprocessed by removal of motion-corrupted volumes using a neural network classifier (see chapter 6) presented in [Kelly et al., 2017], PCA-based denoising [Veraart et al., 2016], distortion and motion correction with outlier replacement [Andersson et al., 2016], bias field correction [Tustison et al., 2010] and intensity normalisation across datasets based on image intensity in high-FA voxels.

8.4.3. Response function estimation

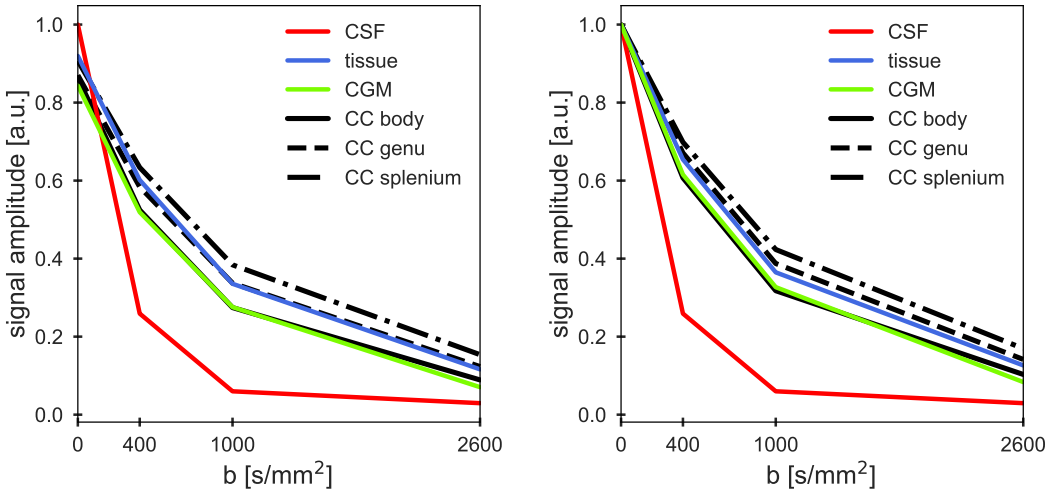


Figure 8.9.: Average signal decay in CSF, single fibre voxels (tissue), cortical grey matter (CGM), and the corpus callosum. left: normalised to the b=0 signal of the CSF component, right: each function normalised to 1. For comparison to the response functions in the adult data, see fig. 8.4.

Subject-specific CSF and WM tissue probability maps, sourced from segmented co-registered T_2-weighted images [Makropoulos et al., 2014], were downsampled to the resolution of the diffusion data. These maps were thresholded at 80% to exclude voxels with partial voluming and used to constrain the single fibre voxel search performed on the b=2600s/mm^2 shell [Tournier, Calamante, Connelly, 2013]. WM responses were subsequently extracted from the resulting single fibre mask. The CSF response function was estimated by selecting the 100 voxels within the thresholded CSF mask with the highest signal attenuation between the averaged b=0 and b=2600s/mm^2 shells. Note that not all white matter tracts were part of the WM single fibre voxel selection due to constraints in the probabilistic masks but the masks prevent the inclusion of GM voxels.

These response functions were then averaged across subjects, and used in the MSMT-CSD decomposition to generate 'free water' density and 'tissue' ODF maps for each

subject. See fig. 8.10 for voxels selected for response function estimation and the resulting decomposition for a single subject. The average signal decay curves are shown in fig. 8.9. To emphasize that the response function derived from WM voxels is not specific to the signal of white matter, it is denoted as 'tissue'.

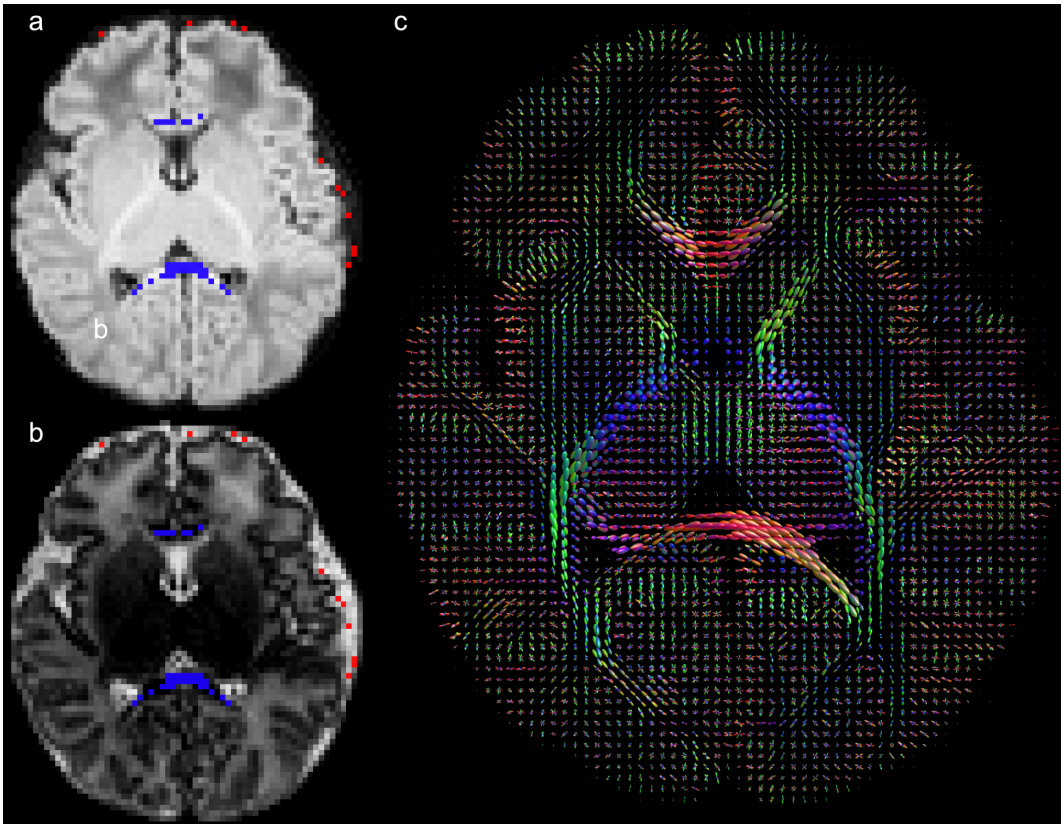


Figure 8.10.: Axial slice from a single dataset with very little motion corruption. Left: voxels selected for the 'tissue' (blue) and 'free water' (red) response function estimation are overlaid onto the 'tissue' density map (a) and onto the 'free water' map (b). Figure c shows the 'tissue' ODF map.

8.4.4. Multi-component template generation

Similarly to the multi-component registration of adult WM and GM, the neonatal 'tissue' and 'free water' maps can be jointly registered. A major difference between both cohorts is that, in neonates, components are not sharply separated but overlap in most areas of the brain (see fig. 8.10). Between the adult and the developing connectome data, the spatial resolution relative to the size of the brain, the distribution of anisotropic signal, and its spatial arrangement differ substantially (compare fig. 8.4). We created two neonatal population templates, one using only the 'tissue' component and the second one using both components combined to validate the effect of using an aggregate metric

combining both HARDI-based components. Prior to registration, dHCP images were up-sampled by a factor of 1.6 to increase the final resolution of the template but the following template-building procedure was identical to that for the adult templates.

As in the adult case, both neonatal templates differ little in their appearance on visual comparison. The most visible difference between both templates is the extent to which superior white matter tracts are aligned (fig. 8.11). The degree of anisotropy, as measured by the square root of the power P_l

$$P_l(I(\mathbf{x})) = \frac{1}{4\pi} \sum_{m=-l}^{l} (I(\mathbf{x}, l, m))^2 \tag{8.12}$$

in the second order spherical harmonics ($\sqrt{P_2}$) is slightly higher in subcortical white matter of the template that used both components for registration (see arrows in fig. 8.11). Therefore, the template that was generated using multi-contrast registration was selected for further discussion in section 8.5.

8.5. Group-level observations in the neonatal template

As expected, compared to a single subject, the template is smoother due to the anatomical and developmental diversity of the cohort (compare figs. 8.4 and 8.10) but the signal decomposition and registration approach proposed here provides good alignment across subjects on visual inspection, with clear definition of features such as the motor strip, brainstem, and anterior commissure figs. 8.13 to 8.16.

Early maturing white matter in the cerebellum, the cerebellar peduncle (fig. 8.14), the CC, and the corticospinal tract (CST) (fig. 8.13) show a high 'tissue' and low 'free water' density. Compared to adjacent areas in the corona radiata, the CST has a low 'free water' and high 'tissue' density and the ODFs are more anisotropic (fig. 8.13), indicating relatively advanced maturation.

The cerebellum and brainstem are relatively mature at birth and have a high cellularity and low free water content (see fig. 2.1, section 2.2.1). This is evident in the low 'free water' density in this area (fig. 8.14) and the structural similarity between the adult and neonatal template. See fig. 8.17 for a direct comparison of an axial slice through the CST, brainstem and cerebellum at the level of the middle cerebellar peduncle. The high angular and spatial resolution of the diffusion data allows a clear and immediate separation of the CST, the middle cerebellar peduncle, the inferior cerebellar peduncle, and of transverse pontine fibers.

Conversely, parts of the anterior periventricular deep white matter exhibit the opposite characteristic (fig. 8.15). This area has a relatively low 'tissue' density pocket and a high 'free water' content, which is consistent with previous observations in diffusion MRI [Judaš et al., 2005].

In the cortex, we see clear radial organisation, consistent with the known process of cortical formation (fig. 8.16). Anisotropy in this area has been shown to drop as dendritic arborisation proceeds [McKinstry et al., 2002]. The arrow in fig. 8.16 a points

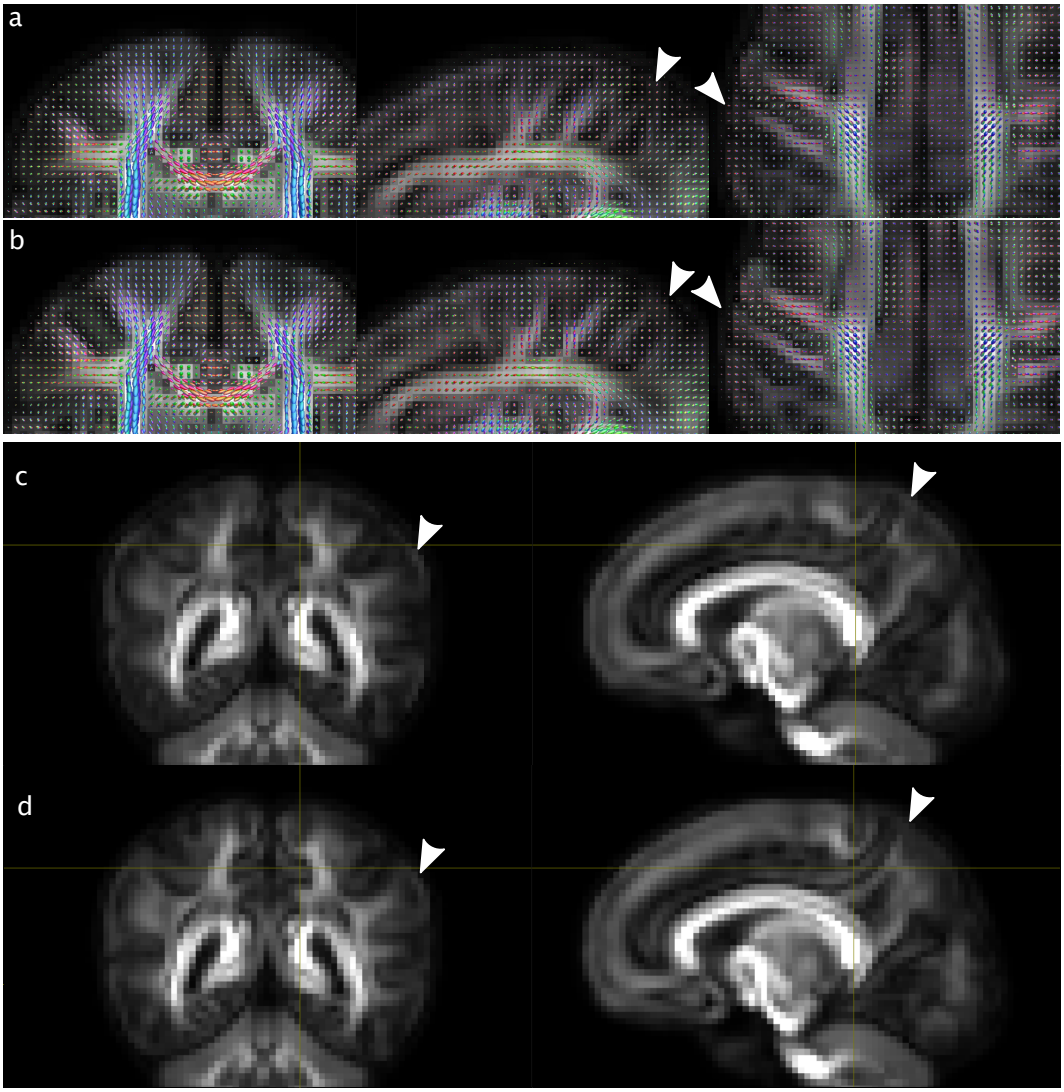


Figure 8.11.: Comparison of the neonatal templates generated using only the 'tissue' component (a,c) and that where registration was driven by the 'tissue' and 'free water' component (b,d). The background image is a measure of anisotropy of the spherical harmonics ($\sqrt{P_2}$).

at the radial organisation of ODFs extending from the WM into the frontal temporal cortex. The ODF anisotropy map in this region shows a low anisotropy between cortical GM and WM as previously shown in high resolution DTI and histology [Miller et al., 2011]. As expected, this is not observed in the gyral crown.

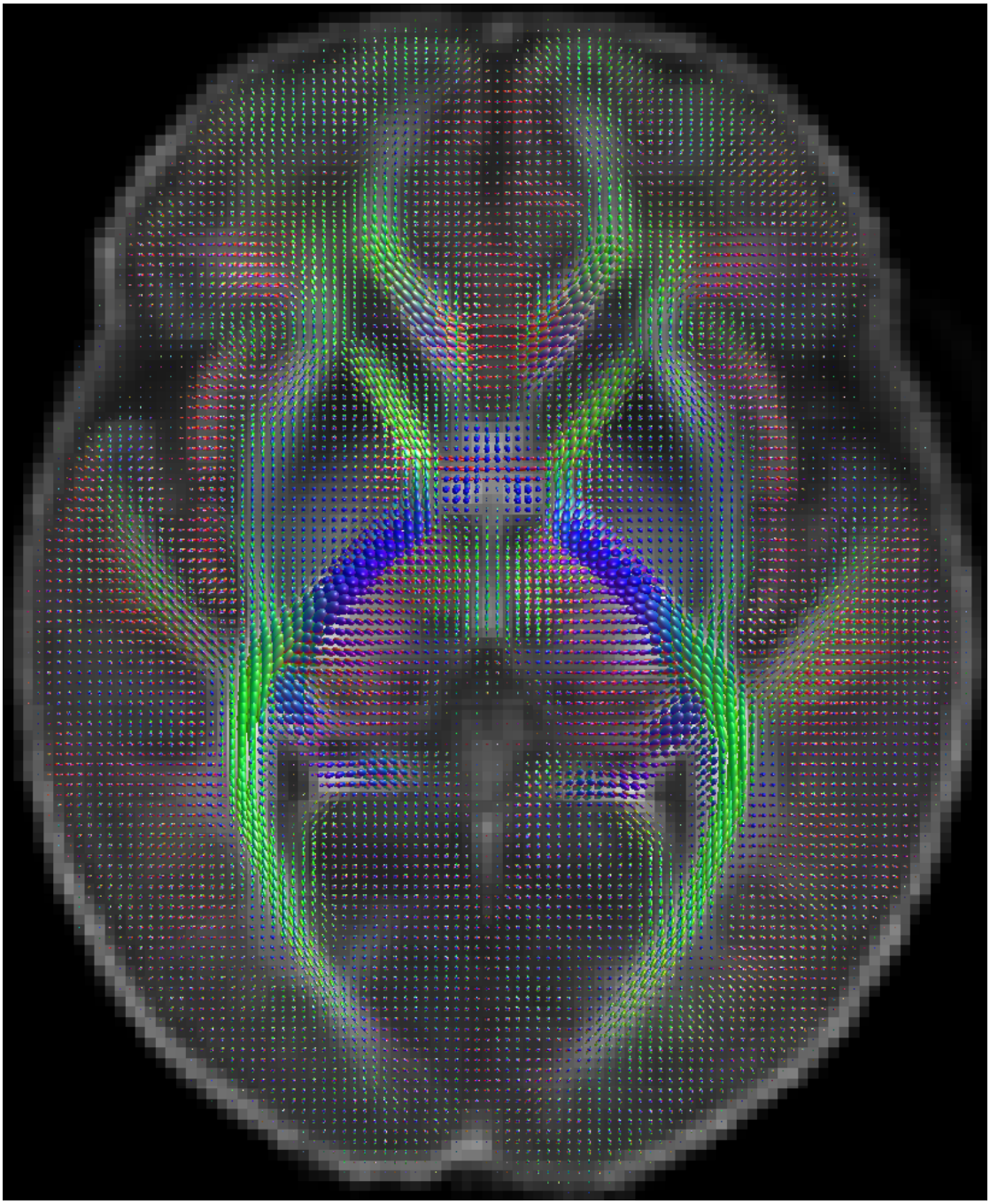


Figure 8.12.: Axial slice of the neonatal template showing the 'tissue' ODFs overlaid onto the 'free water' density image in approximately the same location and orientation as the single subject image in fig. 8.10.

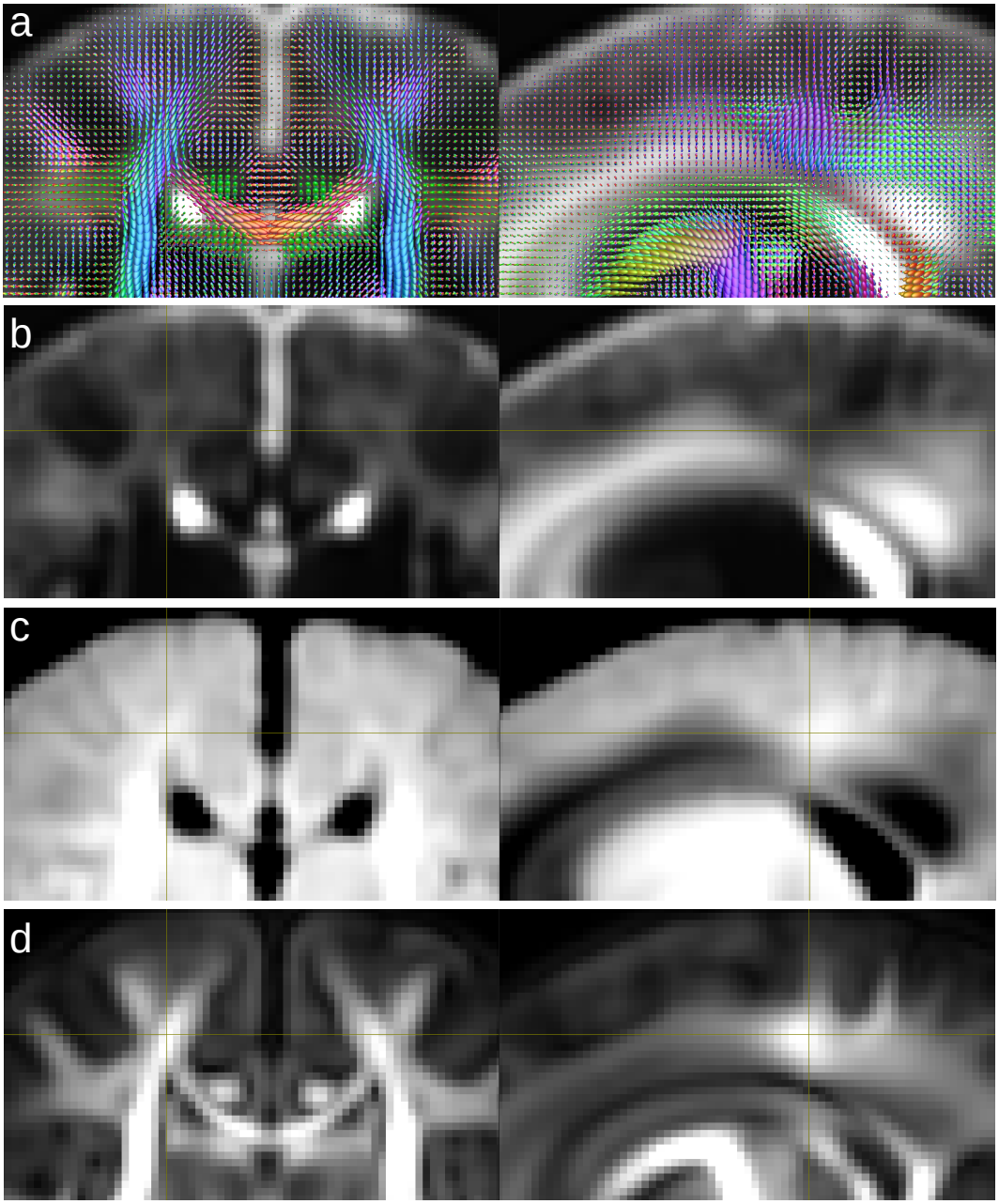


Figure 8.13.: CST and projection fibres. The top row shows 'tissue' ODFs overlaid onto the 'free water' component. b: The CST exhibits a low 'free water' density (cross hair) compared to surrounding white matter of the corona radiata. This, and the high fibre density (c), and the high anisotropy ($\sqrt{P_2}$) (d) suggest more advanced maturation compared to other white matter tracts.

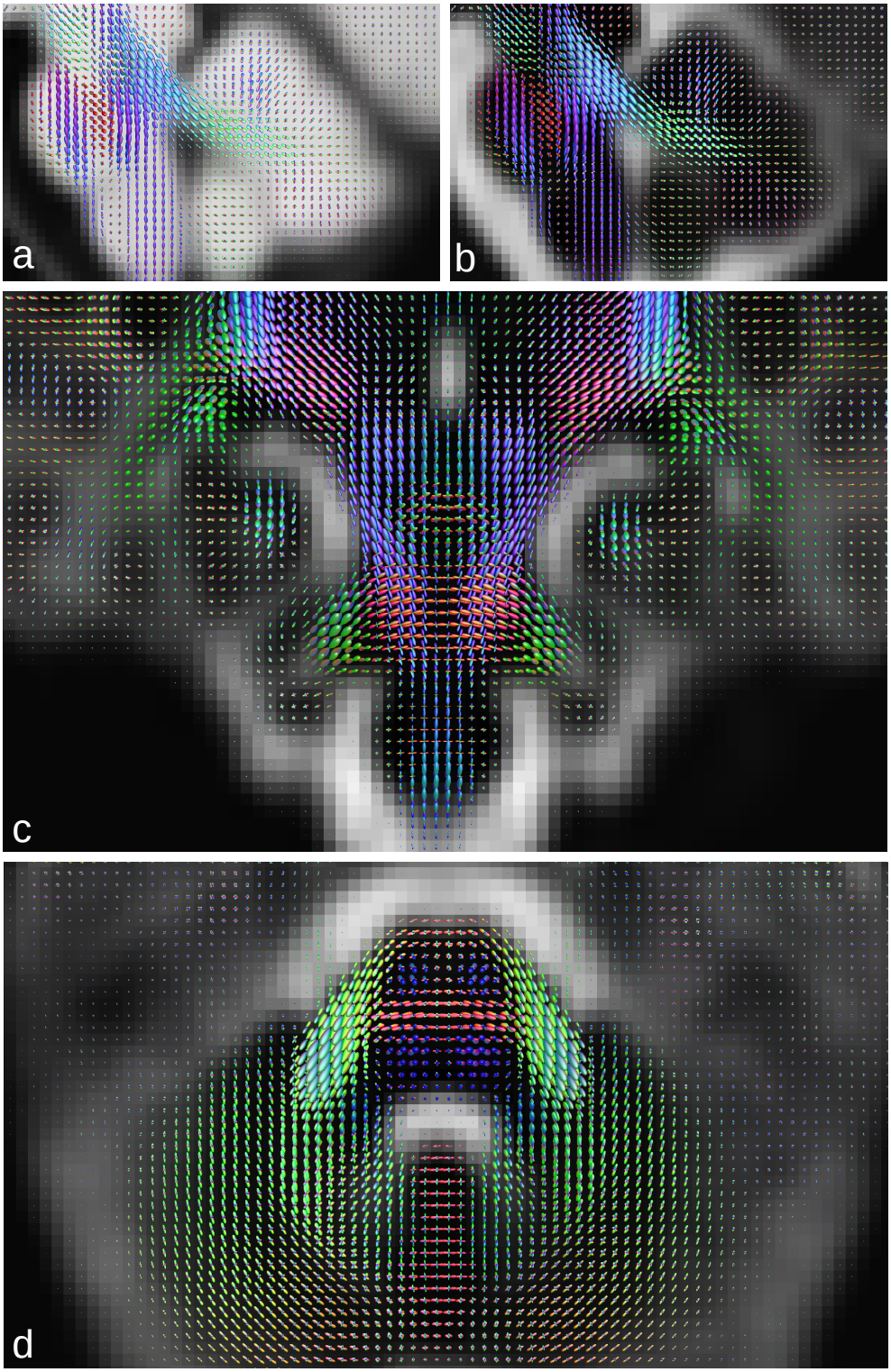


Figure 8.14.: Cerebellum, brainstem and cerebellar peduncles. Sagittal view of cerebellum with ODFs overlaid on 'tissue' density image (a) and 'free water' density image (b). Figure (c) shows a coronal view through the brainstem and the cerebellar peduncles with the 'free water' density map in the background. An axial view through the pons and cerebellum (d) shows the relative maturity of white matter in this region.

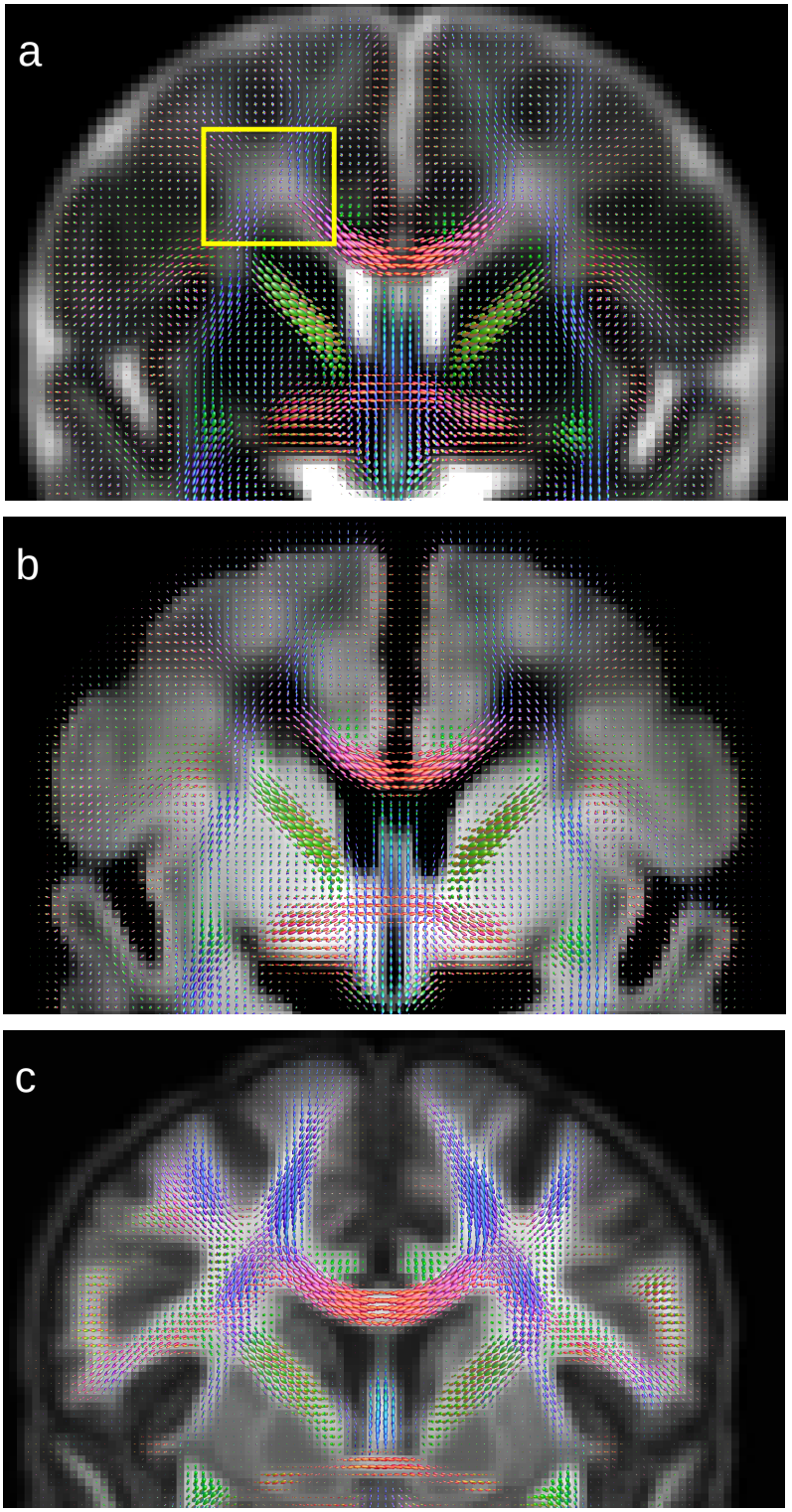


Figure 8.15.: Low 'tissue' density pocket in the area of the frontal periventricular crossroads. Neonates have a high 'free water' content (a, yellow box) and low 'tissue' density (b) in this area. For comparison, figure (c) shows 'tissue' ODFs overlaid onto the average 'tissue' density map of the HCP template, showing high 'tissue' density in that area.

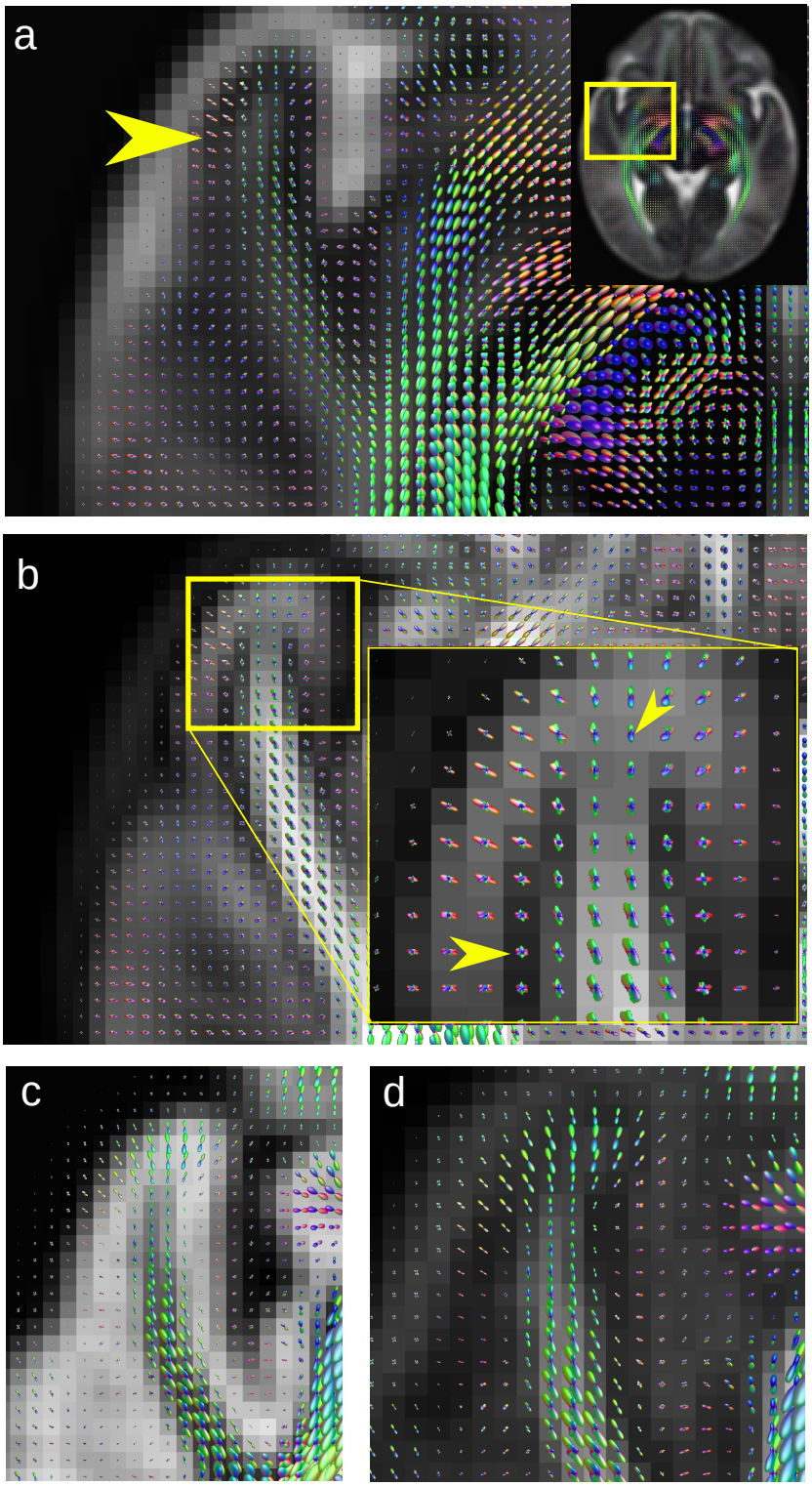


Figure 8.16.: Radial organisation of ODFs extending from the WM into the frontal temporal cortex overlaid on the 'free water' density (a). Neonates exhibit high anisotropy in the cortex likely due to radial glial fibers and pyramidal neurons extending into the cortex [McKinstry et al., 2002]. Image (b) shows the ODF anisotropy in this region. Voxels where high curvature of fibres entering the cortex are expected (b, left arrow) exhibit low anisotropy, in contrast to regions where fibres enter without curving (b, top arrow). For comparison, (c) and (d) show the 'tissue' density in this region of the HCP template.

8.6. Conclusion

The developed methods allowed the creation of a high-quality multi-shell HARDI template of the human brain at the time of birth, in the form of the orientationally resolved 'tissue' density and 'free water' density maps. This framework forms the foundation for advanced longitudinal and group-wise investigations into healthy and abnormal brain maturation.

The template matches the expected anatomy and composition of the developing brain in the neonatal period, with an overall high water content and high anisotropy in the cortex. We chose to use a simple model with two tissue types given the difficulty in distinguishing between WM and GM, and other ongoing developmental processes (e.g. WM neurogenesis and pruning, proliferation, maturation) occurring during this period.

The high 'free water' content in the periventricular cross-roads are challenging for tractography algorithms that regularise spurious connections by stopping tracking if the WM ODF amplitude is below a fixed threshold, or if voxels are affected by partial voluming with CSF. This prevents or hinders tracking of thalamo-cortical fibres, especially in younger subjects. The decomposition into 'free water' and 'tissue' components and the resulting information about relative 'tissue' volume fraction could be used to spatially adjust the tracking parameters.

The microstructural information, especially 'free water' component, are potential biomarkers for disambiguating lesions and bleedings, common findings in prematurely born babies. The next chapter investigates the maturation of brain tissue microstructure on a larger time-frame using an additional HARDI-derived component.

Acknowledgements HCP data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

8.7. Appendix

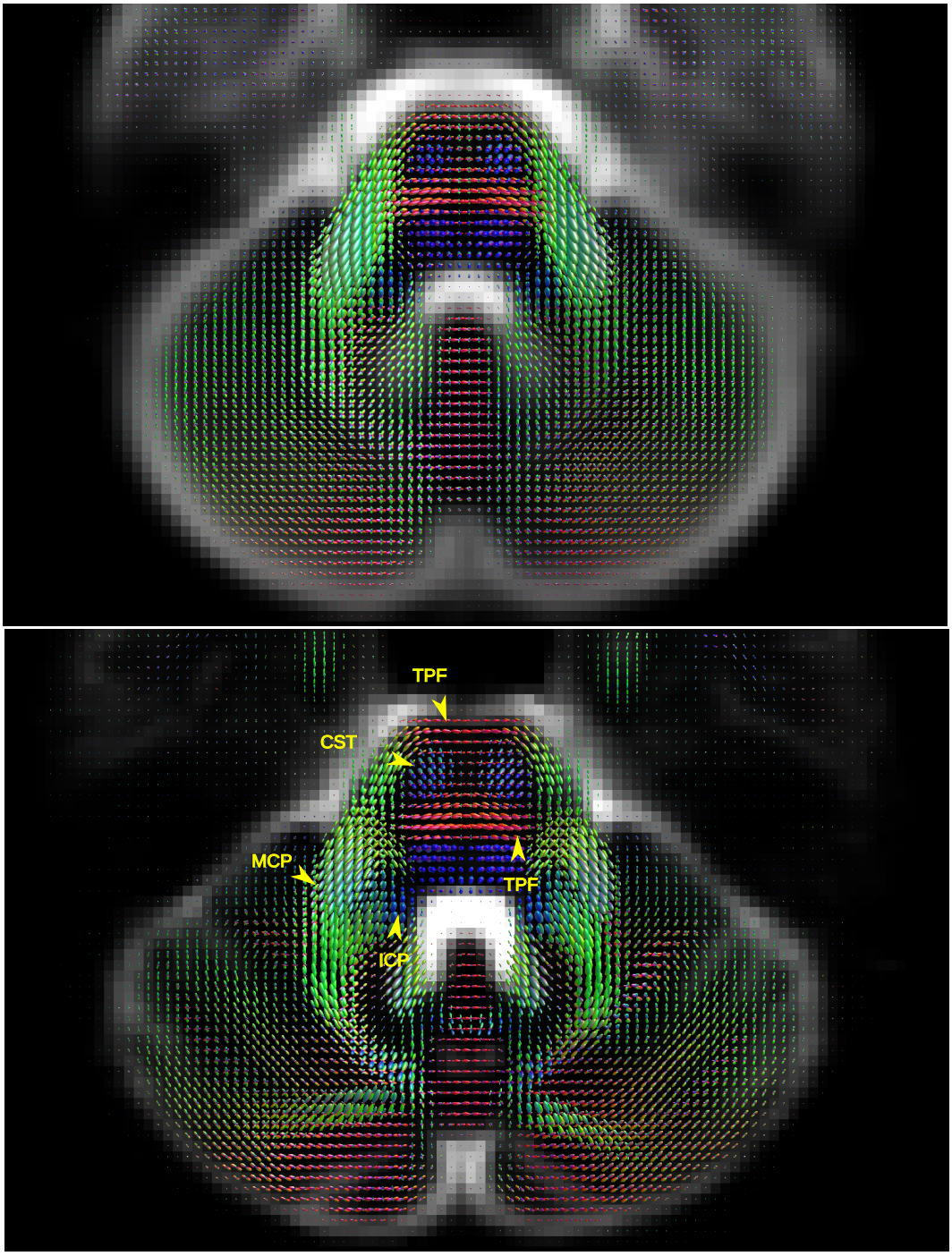


Figure 8.17.: Non-linearly aligned axial slices of the neonatal (top) and the adult template (bottom) showing the organisation of tissue in the cerebellum and the CST. MCP: middle cerebellar peduncle, ICP: inferior cerebellar peduncle, TPF: transverse pontine fibers

Chapter 9

Multi-component HARDI brain atlas over the neonatal period

Contents

We describe a method for creating a time-resolved atlas of the developing white matter using advanced multi-shell high angular resolution diffusion imaging data. This relies on the recently proposed multi-shell multi-tissue constrained spherical deconvolution (MSMT-CSD) technique, and decomposes the signal into one isotropic component and two anisotropic components, with response functions estimated from cerebrospinal fluid, and white matter in the youngest and oldest participant groups respectively. We build a time- and orientationally-resolved atlas of those tissue components from data acquired from 113 babies between 33 and 44 weeks postmenstrual age, imaged as part of the Developing Human Connectome Project. These data were split into weekly groups, and registered to the corresponding group average templates using a previously-proposed non-linear diffeomorphic registration framework, designed to align orientation density functions (ODF). This framework was extended to allow the use of the multiple contrasts provided by the multi-tissue decomposition, and shown to provide superior alignment. Finally, the weekly templates were registered to the same common template to facilitate investigations into the evolution of the different components as a function of age. The final multi-tissue atlas provides insights into brain development and accompanying changes in microstructure, and forms the basis for future investigations into healthy and pathological white matter maturation. A slightly modified version of this chapter has been submitted and accepted, subject to revisions, in the special issue "Imaging baby brain development" of NeuroImage.

9.1. Introduction

Building on the previous chapter, in this work, we describe a method for creating an unbiased atlas of white matter maturation based on advanced diffusion MRI methods, in a cohort of neonates scanned over a range of ages, during which large changes in brain volume, shape and contrast occur. We use high-quality multi-shell High Angular Resolution Diffusion Imaging (HARDI) data acquired as part of the Developing Human Connectome Project (dHCP) to build an atlas of 113 babies scanned just after birth with postmenstrual ages at scan ranging from 32.4 to 44.6 weeks.

The analysis of these microstructural properties in the developing brain requires two main components: a consistent model for the HARDI signal suitable for the neonatal period; and a means of realigning these data onto an unbiased common space. The model used in this work relies on the multi-shell multi-tissue constrained spherical deconvolution (MSMT-CSD) framework [Jeurissen et al., 2014] and requires the determination of appropriate response functions to describe the signal 'signature' for each different tissue component. The image registration is driven based on two such tissue components [Pietsch et al., 2017b], to align subjects within 12 multi-tissue cross-sectional weekly templates, and then jointly to a single time-resolved multi-tissue atlas which is split after alignment into weekly time steps. The atlas itself was created using a decomposition of the signal into three components: one isotropic, derived from CSF, and two anisotropic components, derived from white matter (WM) in the youngest and oldest weekly groups

respectively. The resulting atlas provides a basis for detailed spatio-temporal investigations into healthy and abnormal brain maturation at the single fibre level.

9.2. Materials and Methods

9.2.1. Cohort

The cohort used for this atlas consists of 113 babies scanned as part of the dHCP. From all subjects available, subjects with known clinical abnormalities [Hughes et al., 2017b] and lesions (using Apparent Diffusion Coefficient (ADC), WM or cerebrospinal fluid (CSF) decomposition images) were excluded. If a subject was scanned multiple times, only the first scan was considered. The weekly cohorts have average gestational age at scan of 32.9, 34.0, 35.2, 35.7, 37.1, 38.1, 39.1, 40.1, 40.9, 42.0, 42.8 and 44.1 and consist of 11 subjects, except for the two youngest cohorts and the template at 35.7 weeks which consist of 9, 9 and 10 samples, respectively.

9.2.2. Data

The multi-shell high angular resolution diffusion single-shot spin-echo echo-planar images were acquired on a Philips 3T Achieva scanner using a dedicated neonatal head coil [Hughes et al., 2017a] with a maximum gradient amplitude of 70mT/m. The 300 volumes per image were sampled with four phase-encode directions on four shells with b-values of 0 (n=20), 400 (n=64), 1000 (n=88) and 2600 (n=128) with TE=90, TR=3800ms [Tournier et al., 2015a; Hutter et al., 2017] and reconstructed to a resolution of 1.5mm. The reconstruction method follows the extended SENSE technique proposed in [Zhu et al., 2016]. Sensitivities were estimated from non-accelerated reference acquisitions with matched readouts as in [Hennel et al., 2016] to promote equivalent distortions in the coil maps as in the data.

9.2.3. Preprocessing

The preprocessing of the data consists of: (i) removal of motion-corrupted volumes using a deep neural network classifier (see chapter 6) [Kelly et al., 2017]; (ii) Marchenko-Pastur-PCA-based denoising [Veraart et al., 2016] (MRtrix3); (iii) susceptibility and eddy-current distortion correction and inter-volume motion correction with outlier replacement using *topup* [Andersson, Skare, Ashburner, 2003] (FSL) and *eddy* [Andersson, Sotiropoulos, 2015b] (FSL); and (iv) bias field correction based on the b=0 shell using *N4* [Tustison et al., 2010] (ANTs).

Brain masks were generated using a combination of *bet* [Smith, 2002] (FSL) and a custom-built threshold-based segmentation.

9.2.4. Tissue decomposition

The approach of decomposing the diffusion signal used in this study relies on the MSMT-CSD technique [Jeurissen et al., 2014], which separates the diffusion signal into distinct,

orientationally-resolved tissue types. MSMT-CSD requires multiple, component-specific response functions to be defined, each of which characterises the signal for the corresponding tissue component within each b-value shell, along with its angular dependence. These responses are then used to deconvolve the signal into multiple tissue-specific orientation distribution functions (ODFs).

In adults, the main feature that allows this separation is the fact that different tissue types have sufficiently distinct b-value dependencies, giving a clear separation of the brain into WM, grey matter (GM) and CSF (Figure 9.1). As discussed in the previous chapter, in neonates, this clean separation between WM, GM and CSF does not occur naturally. At term-equivalent age, the average signal in cortical grey matter is nearly indistinguishable from that in the corpus callosum (CC), while most of the peripheral white matter decays much faster with increasing b-value. As shown in Figure 9.1, the variability in mean signal curves between different WM structures is higher than the difference between WM and cortical GM. This makes the separation of GM and WM difficult, but allows the investigation of differences in different WM structures.

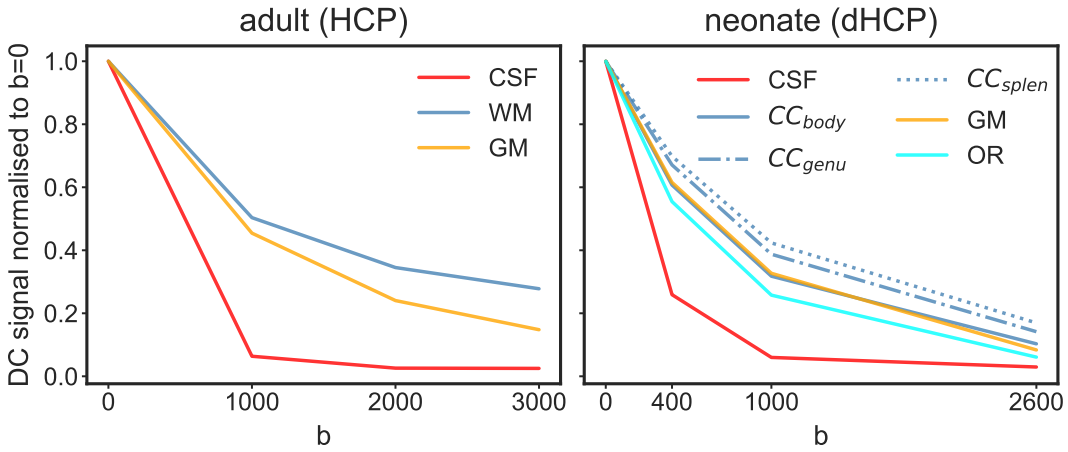


Figure 9.1.: Mean (DC) signal sampled in CSF, WM and GM in adults and neonates at term-equivalent age (40 weeks postmenstrual age (PMA)). In adults, GM and WM are separable by their average signal decay curves. This is not the case for neonates at term-equivalent age where the GM curve lies within the spectrum of WM curves and is very similar to the average signal decay in the body of the CC. Adult data was taken from the Human Connectome Project (HCP). CC_{splen}, CC_{body} and CC_{genu} correspond to the splenium, midbody and genu of the corpus callosum respectively; OR corresponds to the optic radiations.

Furthermore, the WM signal characteristics exhibit a strong age dependence (see section 9.3.1). Therefore, we chose to decompose the diffusion signal using one isotropic component (Iso), derived from CSF voxels, and two anisotropic response functions A_y and A_o. The latter two are derived from dense WM in the youngest and oldest cohorts respectively. The Iso component is equivalent to the 'free water' component in the previous chapter, albeit being derived using a slightly modified procedure described below. This three tissue model serves as a basis to build a time-resolved atlas, where the balance

between the two WM components can be interpreted as an indication of the transition from immature to more mature tissue. This choice is motivated by the observation that the 'WM' response is age-dependent, as shown in section section 9.3.1, and is discussed further in the discussion.

Response function voxel selection The CSF response function was estimated from voxels selected based on their average signal decay within a dilated full brain mask, using the method described in [Dhollander, Raffelt, Connelly, 2016].

The WM response functions were estimated from single-fibre voxels, identified using an iterative procedure described in [Tournier, Calamante, Connelly, 2013]. This was performed with eroded brain masks to ensure only voxels within deep WM were selected. Briefly, the algorithm performs a single-shell CSD using a predefined initial response function; voxels where the main fibre orientation is most dominant are then selected, and an updated response computed by averaging the corresponding DW signal after realignment to a common fibre axis. This process is then repeated until convergence. For younger subjects, brain masks were eroded to exclude most of the cortical GM and single fibre voxel masks were edited manually to remove high Fractional Anisotropy (FA) voxels found in the remaining cortical voxels of younger subjects.

This resulted in consistent WM and CSF voxel selection maps across the age range, as shown in Figure 9.2.

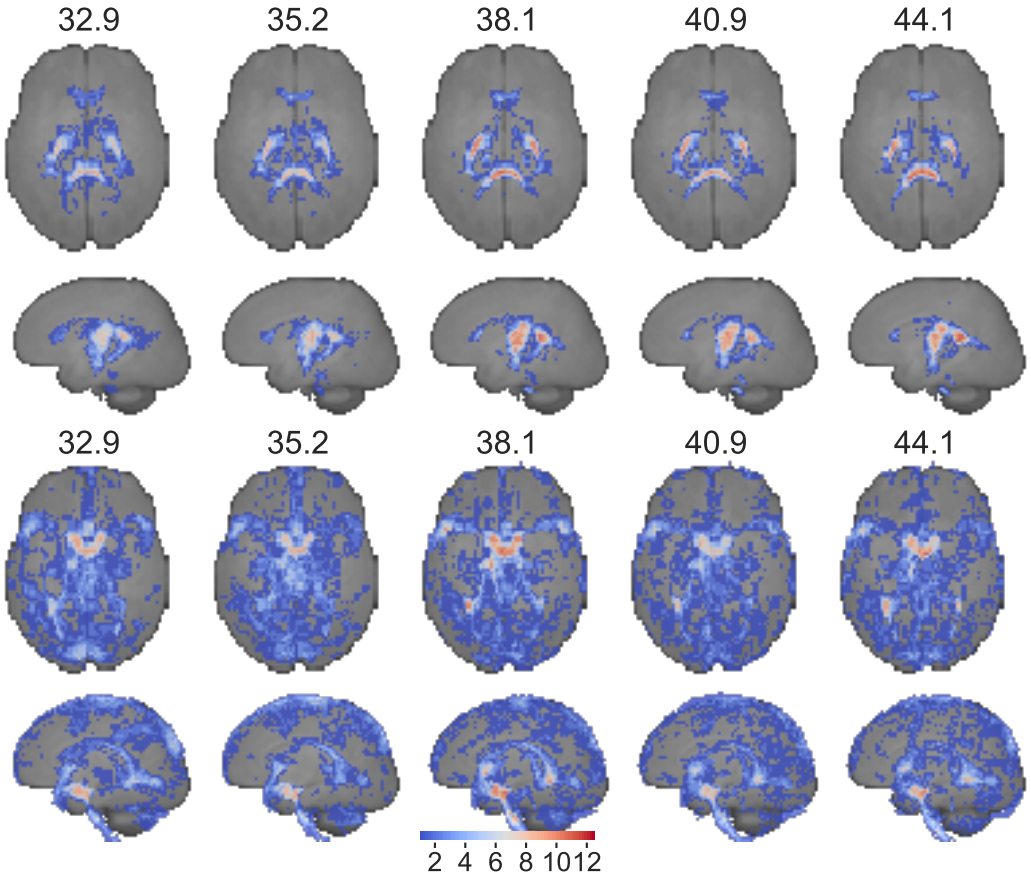


Figure 9.2.: Maximum intensity projection of WM (top) and CSF (bottom) voxel selection masks. Colours represent the frequency a voxel has been selected in the respective weekly template. All images are nonlinearly aligned to a common space.

9.2.5. Bias field correction and intensity normalisation

As previously mentioned, the preprocessing of the data includes a bias field correction step (performed using ITK's N4 algorithm), to minimise any potential influence on the estimated response functions. Any residual bias fields were subsequently corrected following MSMT-CSD by ensuring that the summed density of the three components has an average value of $\sqrt{1/(4\pi)}$; this was performed using the *mtnormalise* command available as part of *MRtrix3* [Raffelt et al., 2017].

9.2.6. Multi-contrast ODF registration

A prerequisite for group-level or longitudinal analysis of HARDI data is unbiased and accurate spatial alignment. We use a symmetric non-linear diffeomorphic registration framework that takes the appropriate ODF reorientation into account [Raffelt et al., 2011] to align individual images to the respective group average image. The registration

cost function metric is the squared L_2 norm of the spherical harmonics coefficients after reorientation between the image and the template, evaluated in the midway-space.

As for the neonatal template, we use the existing ODF registration framework [Raffelt et al., 2011] extended to multiple tissue ODFs to be used simultaneously to drive the registration. In adults, using 'WM' and 'GM' compartments simultaneously (with equal weights) has been shown to to yield higher registration accuracy and sharper features in the spatial and angular domain [Pietsch et al., 2017a]. However, as previously mentioned, a decomposition into distinct 'GM' and 'WM' components is not effective for neonates, prompting us to decompose the tissue into immature and more mature anisotropic components. The problem for registration is that the boundaries between mature and immature tissue are age dependent; using all components would therefore bias the spatial alignment.

For this reason, we decided to use a simpler, two-component decomposition to drive the registration, obtained using responses consisting of each subject's native 'WM' response, and the group average 'CSF' response. Across age groups, the 'CSF' component is consistently located primarily in the ventricles, whereas the single 'WM' component covers the whole brain except for the ventricles and contains the orientational information necessary for the alignment of WM bundles. We find that using the 'CSF' and native 'WM' component for group alignment produces sharper templates (see fig. 8.11) compared to registration using the native 'WM' component only.

We calculated the deformation fields for each subject by registering each subject's 'CSF' component and single native 'WM' component with equal weights to the respective two component version of the template, and subsequently applied those warps to the subject's three component decomposition to create the final three component atlas.

9.2.7. Group average template creation

We created unbiased weekly atlases by iteratively averaging the respective registered images to the corresponding group average template for that week. The templates were created in 28 stages, with increasing degrees of freedom for the transformation and with increasing spatial and angular resolution. More specifically, these stages consisted of six rigid followed by six affine registration stages, each with decreasing voxel sizes from 3.3mm to 1mm and increasing angular resolution ($l_{max} = 2$ for the first four, followed by four with $l_{max} = 4$), followed by 16 nonlinear registration stages. The nonlinear registration increases spatial resolution in eight steps from 3.3mm to 1mm voxel size using $l_{max} = 2$, followed by eight iterations with $l_{max} = 4$ at full spatial resolution. For each iteration in each stage, the update and displacement fields are smoothed using a Gaussian kernel with a standard deviation of 2.0 and 1.0 times the size of the stage's voxel size, respectively. Warps are upsampled using linear interpolation if the resolution is increased between stages. Each image is registered to the group average formed from all other images, *excluding* the current image (leave-one-out) to ensure faster convergence.

Finally, we built a common atlas that aligns all images to a common space to visualise and analyse temporal variability of the three components. The procedure of the joint atlas is identical to the creation of the separately aligned weekly templates. We use the

resulting transformations to build weekly templates that are all aligned to the same space. Note that following the pre-processing, images are interpolated only once to transform them to their respective template space.

9.3. Results

9.3.1. Temporal evolution of the white matter response function

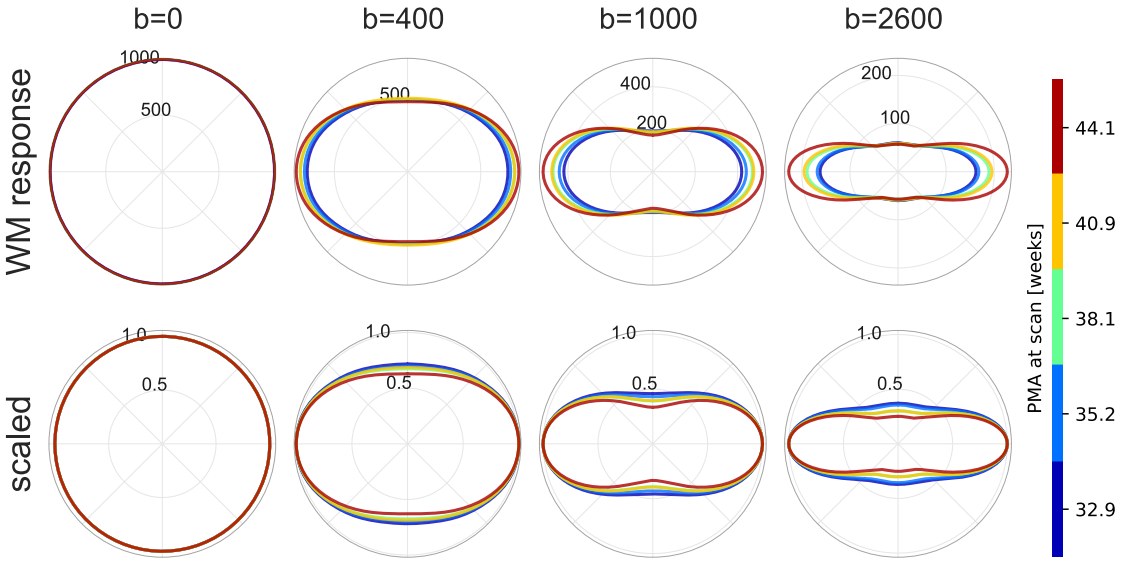


Figure 9.3.: Longitudinal evolution of the WM response function of subjects scanned at 32.9 (n=9), 35.2 (n=11), 38.1 (n=11), 40.9 (n=11) and 44.1 weeks PMA (n=11). Top: Shape and size change visualised as 2D projections through the fibre axis for each shell. Bottom: Each response function scaled independently at each b-value to unit radius to visualise the change in shape.

The WM response function estimated from each subject individually shows a clear age trend in the DC ($l = 0$) and the higher order harmonic coefficients between 32.9 and 44.1 weeks PMA (figure 9.3). With increasing age, the WM response function increases in sharpness and the signal decay across b-values reduces. Note that these weekly response functions exhibit distinct b-value dependencies and are not scaled versions of a single response function. This suggests that, besides CSF, at least two components are required to model the WM signal in neonates accurately, and motivates the use of two anisotropic response functions as was performed in this study.

9.3.2. Multi-tissue model component selection

Given the approximately linear temporal evolution of the WM response functions as a function of age, we postulate that WM maturation can be modelled as a weighted sum of two responses. We use the WM response function of the youngest and oldest age

group, A_y and A_o respectively, to test whether we can express the WM response function $\mathrm{WM}(t)$ at any age t as a linear combination of those response functions by solving the least squares problem:

$$\underset{\alpha,\,\beta}{\arg\min} \left(\alpha * A_y + \beta * A_o - \mathrm{WM}(t) \right)^2 \tag{9.1}$$

Figure 9.4 shows that any response function in the cohort is well represented by a positive weighted sum of the two average response functions of the two age extremes of the cohort. The relative weight between the response functions transitions smoothly from the youngest to the oldest group, suggesting that MSMT-CSD performed using these responses can give meaningful separation of maturation patterns in WM, with the balance of density between the estimated weights for the two responses representing the level of WM maturation.

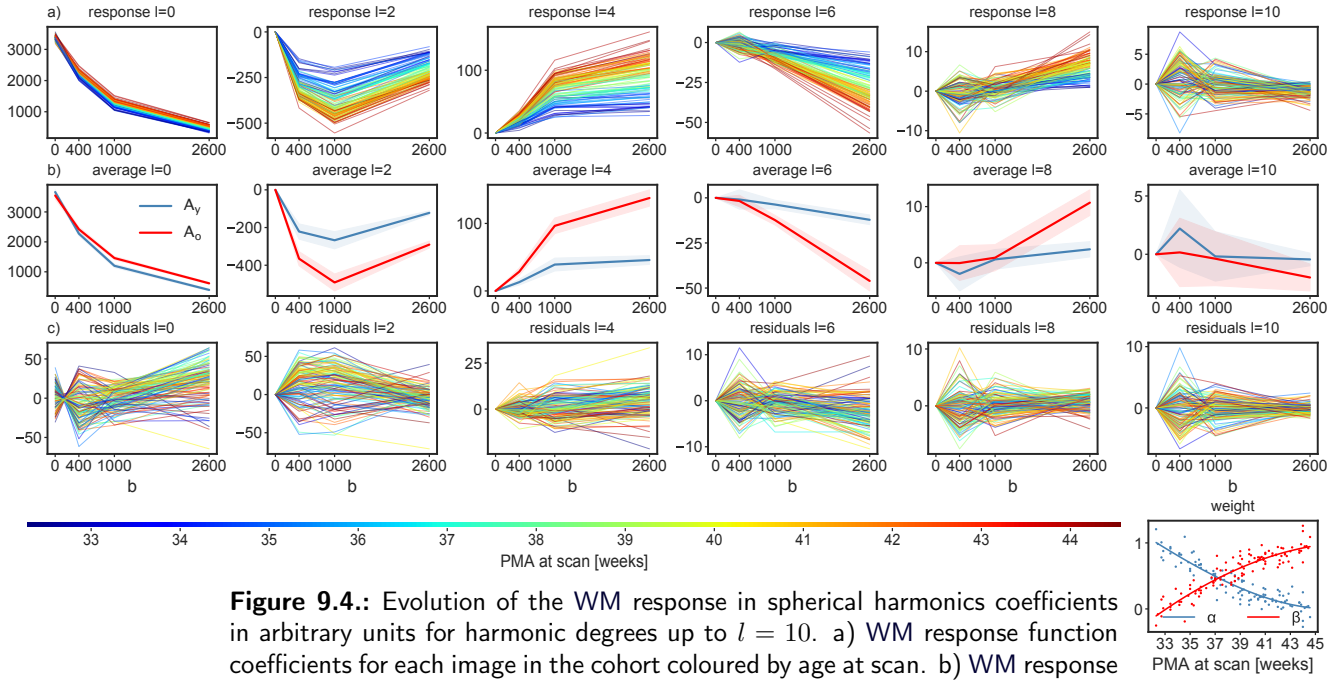


Figure 9.4.: Evolution of the WM response in spherical harmonics coefficients in arbitrary units for harmonic degrees up to $l = 10$. a) WM response function coefficients for each image in the cohort coloured by age at scan. b) WM response function average and 68% confidence interval of 9 neonates scanned at 32.9 weeks (A_y) and of 11 neonates scanned at 44.1 weeks PMA (A_o). The plots in row c) show the residuals (fit - data) and weights (α and β) of the linear model fit defined in equation 9.1. Weights are not constrained to be non-negative. The two curves for α and β are cubic polynomials fitted using a Huber kernel.

The appropriateness of the model was also investigated by looking at the residuals of each MSMT-CSD fit across the age range, as shown in figure 9.5. Using the three component decomposition yields lower residuals than any of the two-component models, which included the CSF (Iso) and a single WM response function (including notably the case where each subject's native WM response function is used). However, there nevertheless remains anatomical structure in the residuals for the three component decomposition,

indicating that the data may contain further information that is not captured by our model.

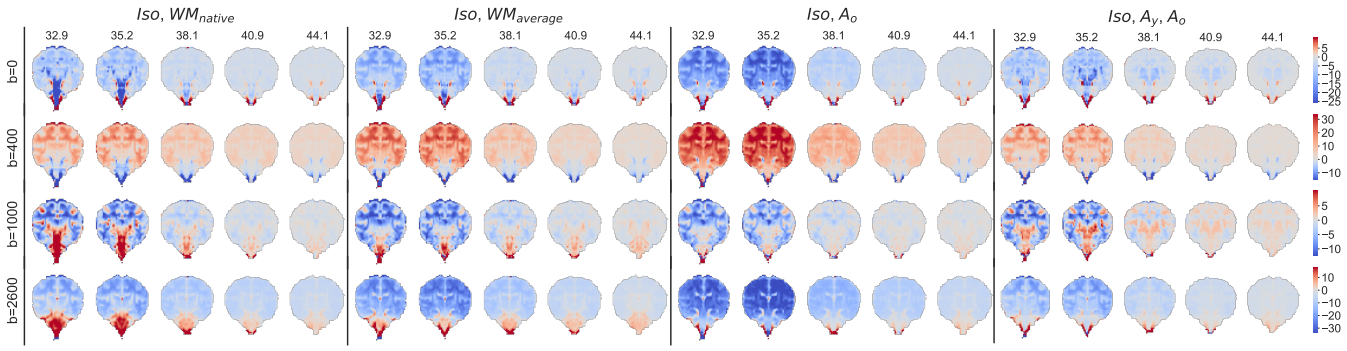


Figure 9.5.: Residuals of the average signal in each shell for different response function combinations. Values are in percent relative to the average b=0 signal. Iso refers to the cohort average CSF response function, WM_{native} to the subject's WM response function, $\text{WM}_{average}$ to the cohort average WM response function and A_y and A_o to the average WM response function of the youngest and oldest cohort, respectively.

9.4. Discussion

9.4.1. Cohort

We chose the subjects from the healthy appearing dHCP cohort to include in this work so that the weekly templates are as unbiased towards gender, age since birth and anatomy as possible. To ensure comparable anatomical variability across age, we selected 9 to 11 subjects per weekly template. If more datasets were available for a template, we ranked them using the following criteria and chose only the 11 best samples.

The images were grouped so that the number of motion artefact free volumes per subject is maximised while minimising both the deviation from normal (age and gender-matched) birth weight and the age since birth. To better balance age range and gender-bias, subjects were assigned to up to two time points. See figures 9.13 for plots of cohort age, weight and quality measures.

9.4.2. Obtaining quantitative density values

In MSMT-CSD, the mapping from diffusion weighted (DW) signal to orientation distribution function (ODF) amplitude is linear and the ODFs obtained are not inherently normalised, which necessitates the use of bespoke normalisation and bias field correction procedures to provide quantitative density values that can meaningfully be combined into a single analysis, such as this atlas. It also requires the use of a single set of responses that are appropriate over the entire age range, so that density estimates can be compared like for like.

First, we estimated CSF and WM response functions for each subject independently, following an initial coarse bias field correction. Those are used to deconvolve the signal into two components which are then subsequently corrected using a more fine-grained bias field correction using *mtnormalise* (*MRtrix3*). The two ODF images are jointly scaled so that they sum on average to $\sqrt{1/(4\pi)}$ within the subject's brain mask. For any further analysis, the subject's response function is also normalised using the inverse of the scale factor applied in the final bias correction. This ensures that the estimated response functions are comparable across age-groups (figure 9.1) but is not strictly necessary for the creation of the atlas.

Following the normalisation of the two subject-specific response functions, we average the normalised CSF response functions of all subjects and the normalised WM response functions of the youngest and oldest cohort and use those to deconvolve the initial bias field corrected DW images. We therefore use the same three response functions for all subjects. The final three component decomposition is bias-field corrected and normalised using *mtnormalise*. This ensures that ODF amplitudes are comparable across all subjects irrespective of their age.

9.4.3. Group-level observations

The component volume fraction maps in figure 9.6 show the decreasing Iso content and the increase of the mature tissue component in brain parenchyma over time. This matches the expected decrease in overall brain tissue water content during development [Dobbing, Sands, 1973]. In our decomposition, early maturing WM such as the cerebellum, cerebellar peduncle or corticospinal tract (CST) exhibit high A_y and A_o and low Iso density at all ages (see figure 9.9). In contrast, parts of the periventricular deep white matter show relatively high free water content (see figure 9.7).

In general, the transition from young to mature-appearing WM occurs from central to peripheral, inferior to superior and posterior to anterior. The slices shown in figure 9.6 display this pattern most prominently in the periventricular cross-roads, the CST and the CC. The sagittal images exhibit a pattern of transition from young to more mature appearing WM in the cerebellum and the CC, consistent with the expected behaviour [Branson, 2013].

Figure 9.9 shows the spatially localised A_o component in the youngest cohort. At 32.1 weeks PMA, it is confined nearly exclusively to the genu and splenium of the CC and WM in the CST, the spine, parts of the midbrain and the cerebellum. The transition from A_y to A_o is similar for the genu and splenium of the CC but the body of the CC has a comparatively high density of A_y fibres even at 44.1 weeks PMA.

In the cortex of younger subjects, we observe clear radial organisation (figure 9.7) which reduces with age, consistent with the known process of cortical formation. Anisotropy in this area has been shown to drop as dendritic arborisation proceeds [Miller et al., 2011; McKinstry et al., 2002].

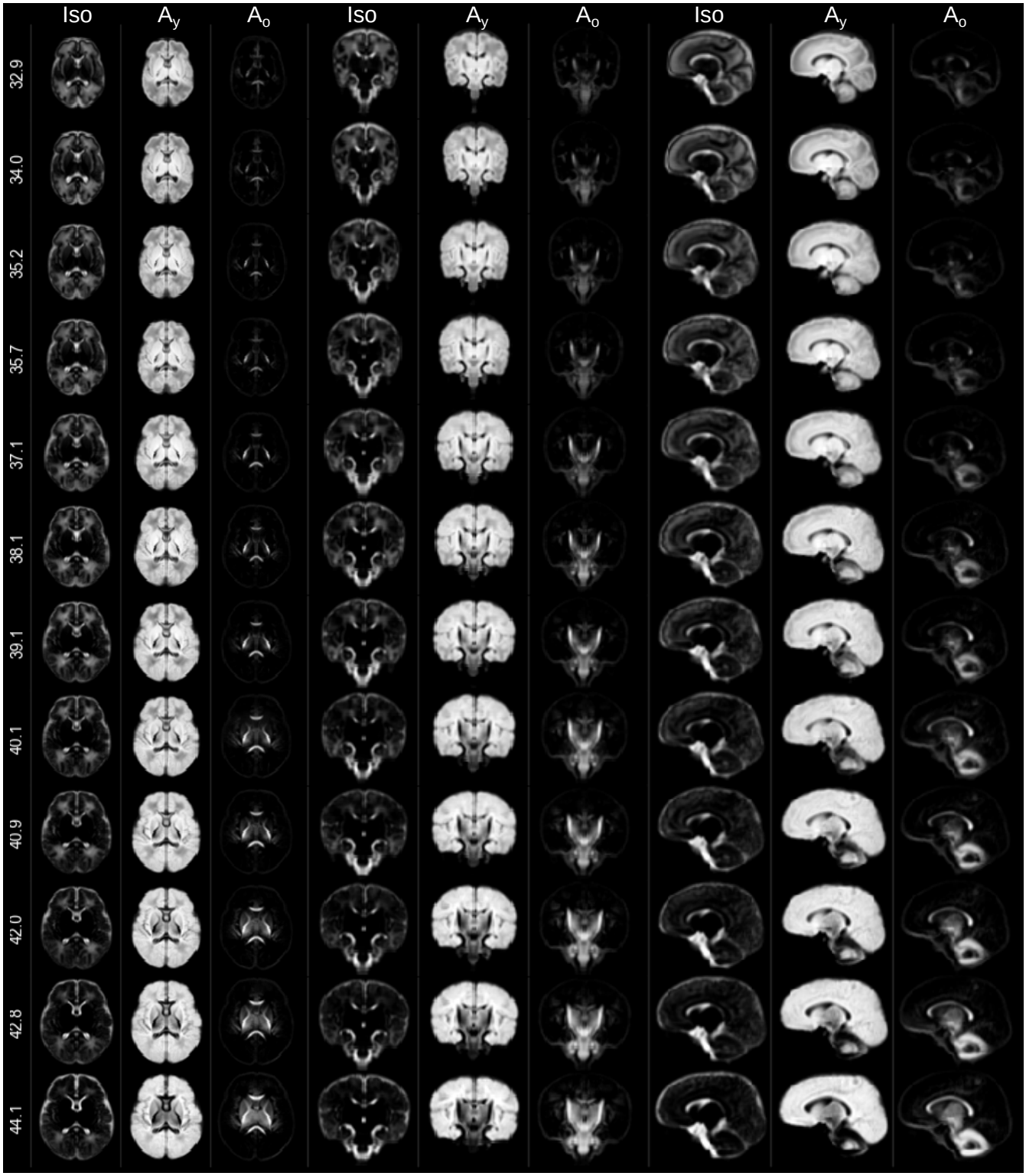


Figure 9.6.: Display of changes in component volume fractions in weekly steps with image intensities representing average ODF amplitude, scaled identically across components and weeks. Note that different anatomical orientations are scaled differently in size.

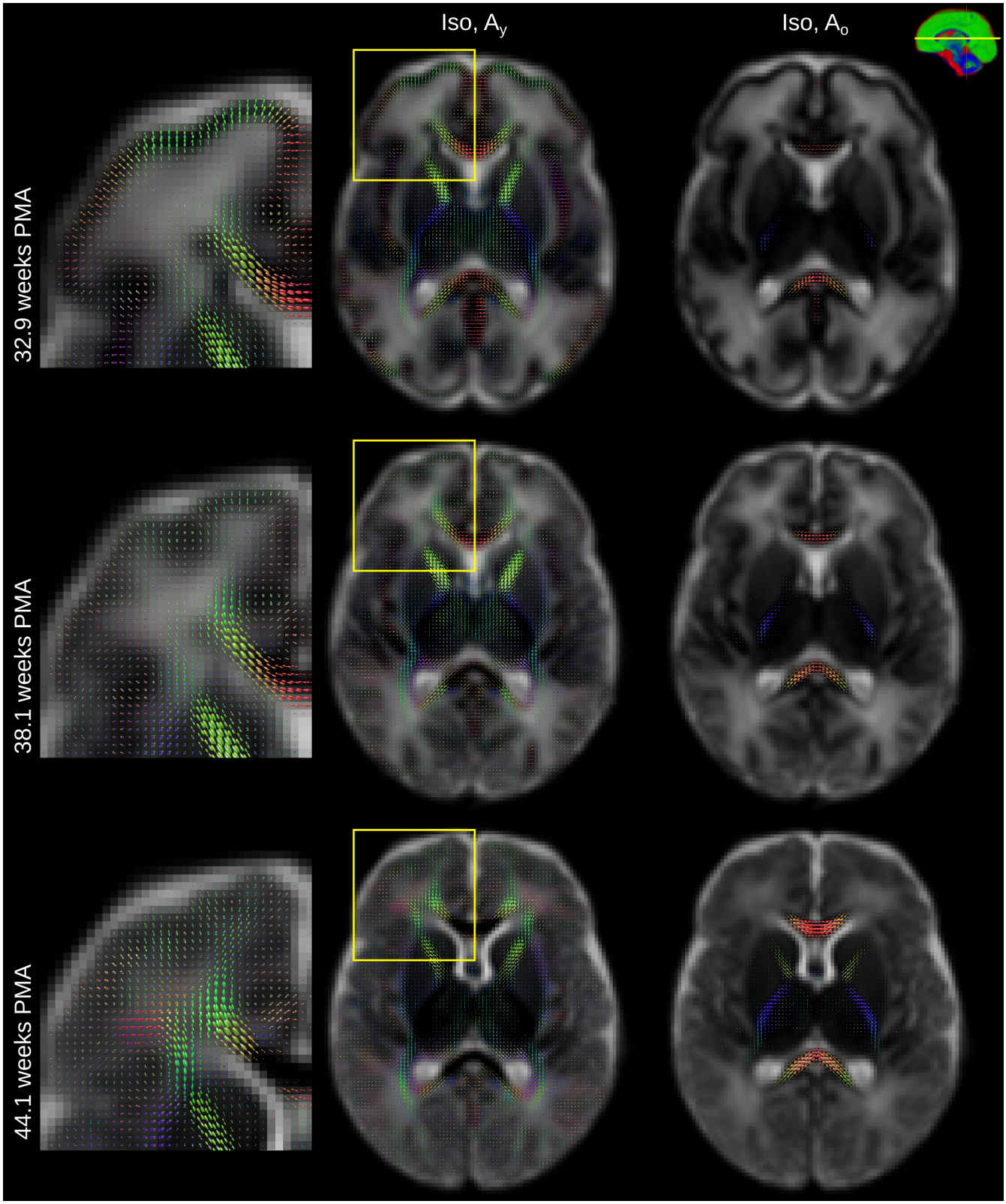


Figure 9.7.: Axial sections showing the isotropic component (background) and the two anisotropic components through the CC and periventricular cross roads at 32.9 (top row) and at 44.1 (bottom row) weeks PMA. Magnified cropped images show high anisotropy in the the cortical GM and high Iso component volume fraction in the periventricular cross roads observed in the young subjects. All images are part of the jointly aligned atlas.

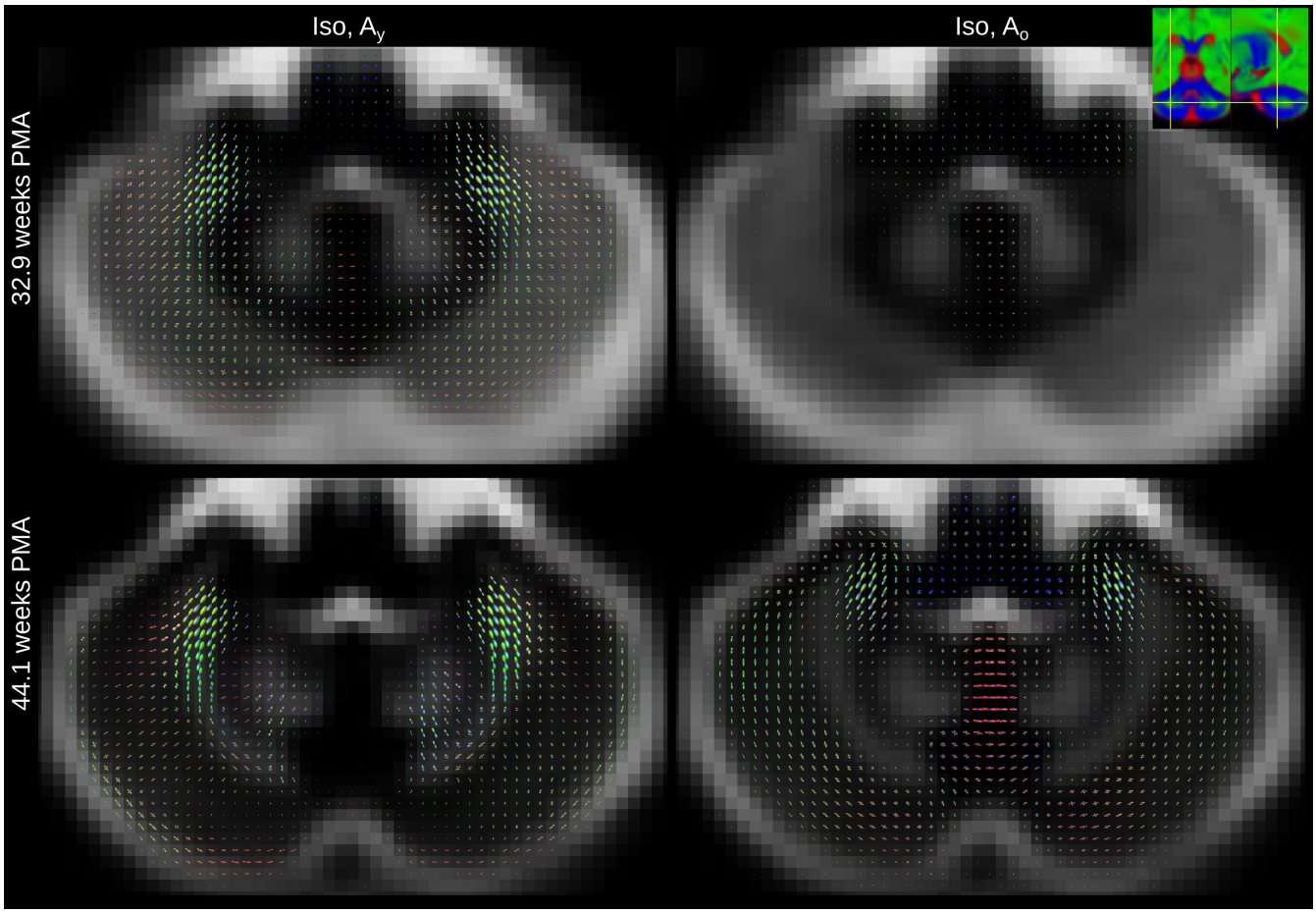


Figure 9.8.: Axial sections showing the isotropic component (background) and the two anisotropic components through the cerebellar dentate nucleus at 32.9 (top row) and at 44.1 (bottom row) weeks PMA. Note the orthogonal opposed fibre directions of the A_y and A_o ODFs in the lateral cerebellar hemispheres. All images are part of the jointly aligned atlas.

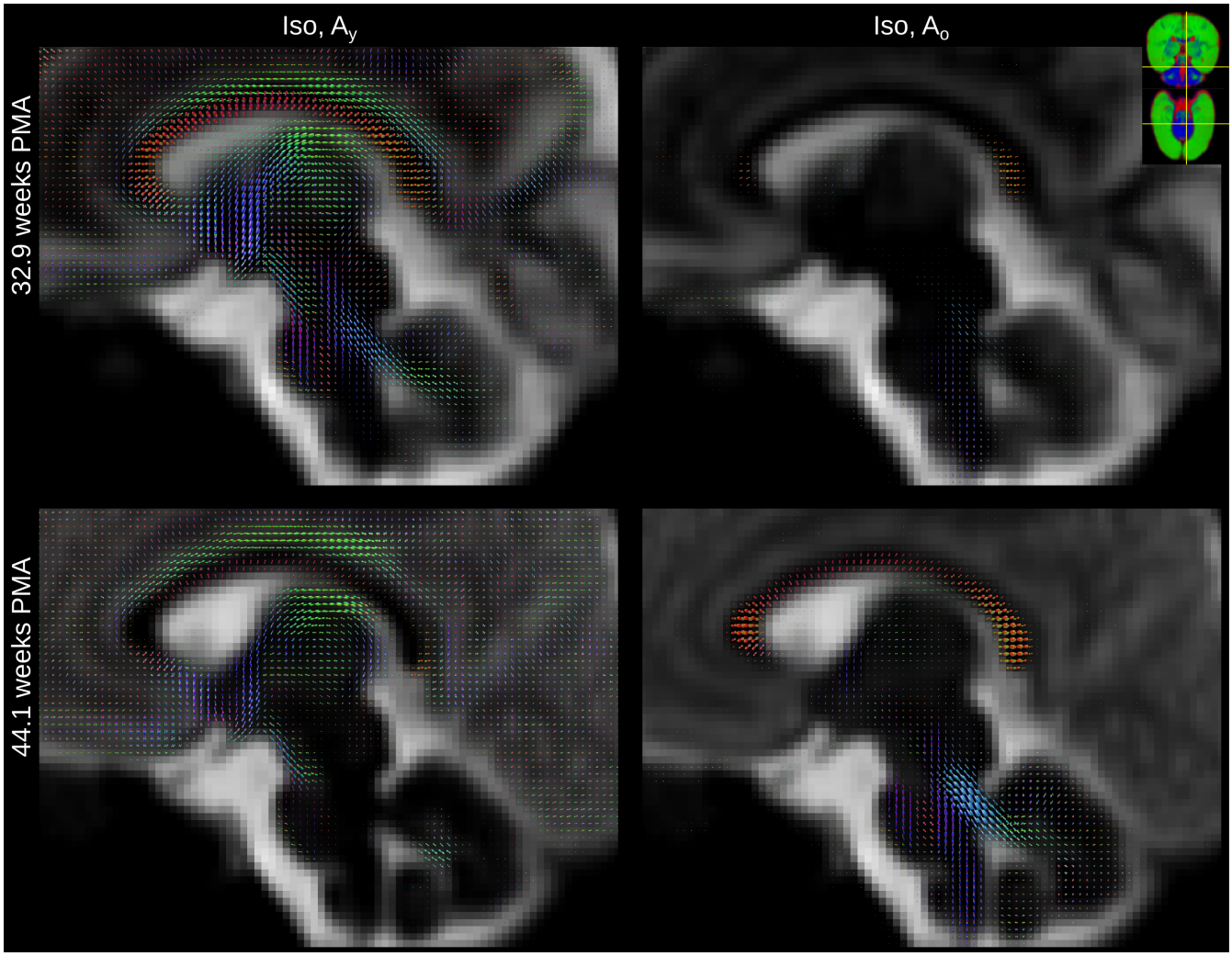


Figure 9.9.: Sagittal sections showing the isotropic component (background) and the two anisotropic components through the brainstem at 32.9 (top row) and at 44.1 (bottom row) weeks PMA. All images are part of the jointly aligned atlas.

9.4.4. Time-resolved component volume fraction changes in selected regions

Histology studies have reported spatially varying onset and progression of myelination [Brody et al., 1987; Gilles, Shankle, Dooling, 1983]. Myelination progresses in a nonlinear and location-specific manner starting at the end of the fourth fetal month lasting until adulthood in the CC [Kinney et al., 1988]. However, myelinogenesis is preceded by complex changes in cellular constituents and their organisation in the premyelinating stages [Back et al., 2002; Wimberger et al., 1995].

We investigate regional differences of WM maturation patterns similarly to earlier studies which used DW imaging [Bui et al., 2006], diffusion tensor imaging [Zanin et al., 2011] or HARDI [Kunz et al., 2014]. Our work differs from this early work in that we use tissue-specific responses instead of biophysical model quantities. Furthermore, in contrast to tensor-based work, our approach can resolve multiple fibre populations within a single voxel. In fact, in some crossing fibre regions, the different bundles are ascribed to different anisotropic responses, potentially reflecting different stages of maturation for the different bundles (figure 9.12).

We manually segmented 17 white and grey matter structures in the jointly aligned group average template using the weekly resolved FA, CSF and both WM component maps. The areas included five regions in the CC from the genu to the splenium, along with further regions in the posterior and anterior limb of the internal capsule, the cingulum, external capsule, fornix, head of the caudate, middle cerebellar peduncles, optic radiation, putamen, superior cerebellar peduncles, thalamus and cortical GM (see figure 9.10).

The middle and superior cerebellar peduncles exhibit a relatively high fraction of A_o at 33 weeks PMA which increases almost linearly until term (superior) and 44 weeks (middle). Of interest, the superior cerebellar peduncle has a higher fraction of A_o from 33 weeks which is consistent with the earlier maturation of the superior cerebellar peduncle compared to the middle cerebellar peduncle [Gilles, Shankle, Dooling, 1983].

We observe that in the posterior limb of the internal capsule, the relative fraction of the A_o component increases rapidly from 33 weeks until 40 weeks after which it slowly increases until 44 weeks. This is in contrast to the anterior limb of the internal capsule which starts to transition from A_y to A_o only after 39 weeks. This is in agreement with reported temporal maturation time courses for these two adjacent structures, as myelin is present in histological sections of the posterior limb of the internal capsule starting at 34 weeks, and myelinates rapidly until after term [Gilles, Shankle, Dooling, 1983], whereas the anterior limb of the interior capsule shows no evidence of myelination until after term [Gilles, Shankle, Dooling, 1983].

In comparison, the external capsule, the fornix, the cingulum and the optic radiations do not myelinate before term [Gilles, Shankle, Dooling, 1983; Yakovlev, Lecours, 1967]. We also observe a later onset of increasing A_o volume fractions in those structures.

In the CC, the splenium appears to mature before the genu and the body exhibits a more protracted maturational pattern. The splenium has been observed to mature before the genu on T1- and T2-weighted imaging [Barkovich et al., 1988].

The pattern of maturation in deep GM is distinct from that of early maturing WM.

The head of the caudate nucleus, the putamen and the thalamus contain very little A_o signal until 37 weeks PMA. The mature component rises more steeply in the thalamus which matches observed myelogenesis in the last trimester [Yakovlev, Lecours, 1967].

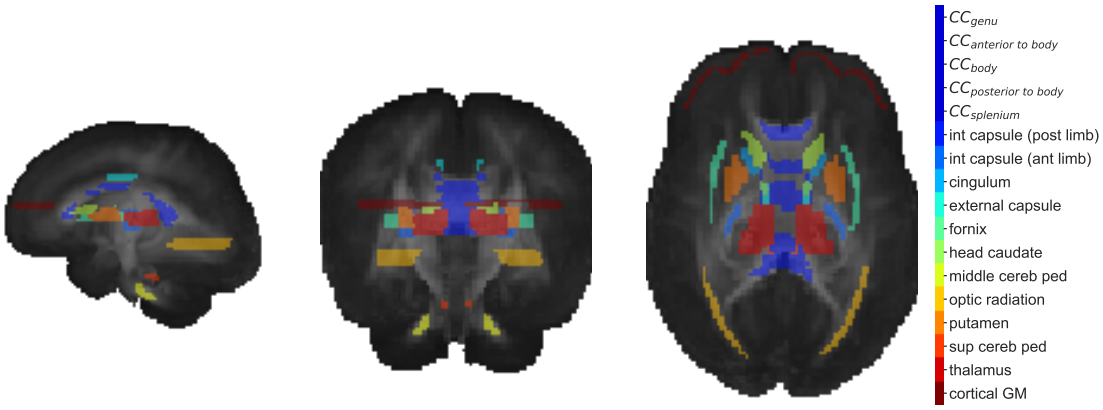


Figure 9.10.: Maximum intensity projections of regions of interest overlaid onto a maximum intensity projection of the age-average FA image. For images of individual ROIs see figure 9.14.

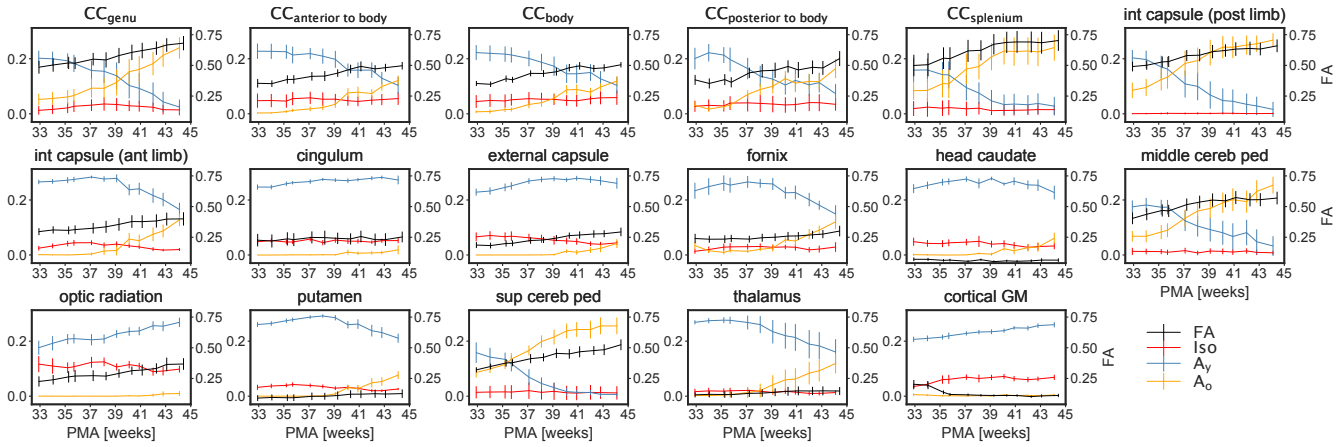


Figure 9.11.: Longitudinal changes of component volume fractions and FA in selected WM and GM regions of interest. Error bars represent one standard deviation across voxels in the respective region and are calculated in the average space.

9.4.5. Limitation of the three tissue model

We focus on modelling the spatial variability of temporal changes in WM. The components of our atlas were chosen to be interpretable in terms of brain maturation. However, WM maturation is undeniably a complex biological process, giving rise to dMRI signals

that might not necessarily be fully characterised using three components alone. Nonetheless, our approach is likely to provide a good first-order approximation to the dominant effects observed in the data over this age range.

It is important to note that the anisotropic WM responses used in this study correspond to the extremes of the age range under consideration, and are therefore inherently dependent on these ages.

Also, the use of CSF (Iso) and two WM response functions is not necessarily applicable to the rest of the brain parenchyma. This is apparent in the MSMT-CSD residual maps of the younger cohorts (figure 9.5). Yet, this work is the first study using a data-driven approach to describe WM maturation during the perinatal period in a fibre-resolved manner. Improving on the response functions selection to model the full brain is scope for future work.

9.4.6. Multiple fibre specific maturation patterns in a voxel

Differentiating between distinct fibre populations within a single voxel based on differences in microstructural features is an ongoing challenge in DW imaging. Microstructure-informed tractography methods have been proposed to disentangle multiple fibre populations [Daducci et al., 2015; Sherbondy, Rowe, Alexander, 2010; De Santis et al., 2016].

MSMT-CSD allows resolving multiple tissue types in the same voxel. Using two anisotropic response functions, we can directly resolve fibre populations from different components in the same voxel if the fibre populations are separable using the chosen response functions. We observe this for instance in the cerebellum. Our three component model separates fibres in cerebellar GM that follow a radial trajectory from tangential fibres within the same voxel (figure 9.8). This matches with observations of radial and tangential pathways [Takahashi et al., 2014] that mature at different rates in the cerebellum.

Furthermore, figure 9.12 shows a section through the CST and the midbrain which illustrates that the "maturation" trajectory for fibres going in inferior-superior direction is distinct from that of pontocerebellar fibers in the same voxel. Resolving multiple maturation patterns in a voxel opens new possibilities for time-resolved investigations in a fibre-specific manner using frameworks such as fixel-based analysis [Raffelt et al., 2016]. Note that this ability to resolve different fibre populations based on their distinct microstructural signature is possible due to the large differences that brain development introduces in their dMRI signature; differences of such magnitude are unlikely to be observed in adult data.

9.5. Conclusions

We propose a method to create a time- and orientationally-resolved multi-tissue atlas of the neonatal brain using three components derived from CSF, WM at 32.9 and WM at 44.1 weeks postmenstrual age. We find regionally-varying temporal patterns in the transition between the young and more mature anisotropic components and were able

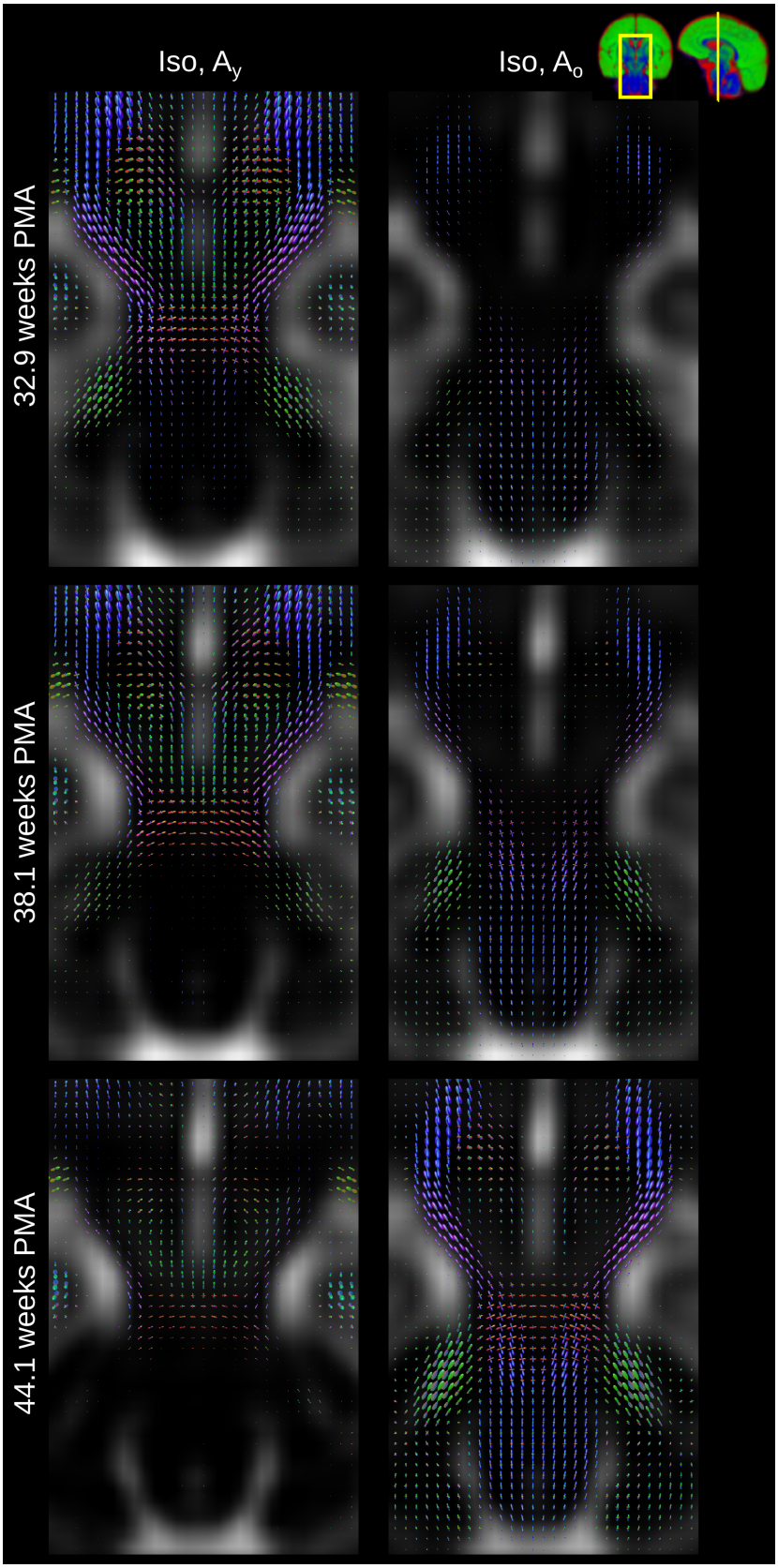


Figure 9.12.: Coronal sections showing the isotropic component (background) and the two anisotropic components through the CST and brainstem at 32.9 (top), 38.1 (middle) and at 44.1 (bottom) weeks PMA. Note the transition from A_y to A_o is different for different fibre populations within the same voxel. All images are part of the jointly aligned atlas.

to distinguish fibre populations within the same voxel with distinct time courses. This atlas provides a basis for investigations into healthy and pathological brain maturation.

Acknowledgments

This work received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 319456, and was supported by the Wellcome EPSRC Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z), and by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

9.6. Appendix

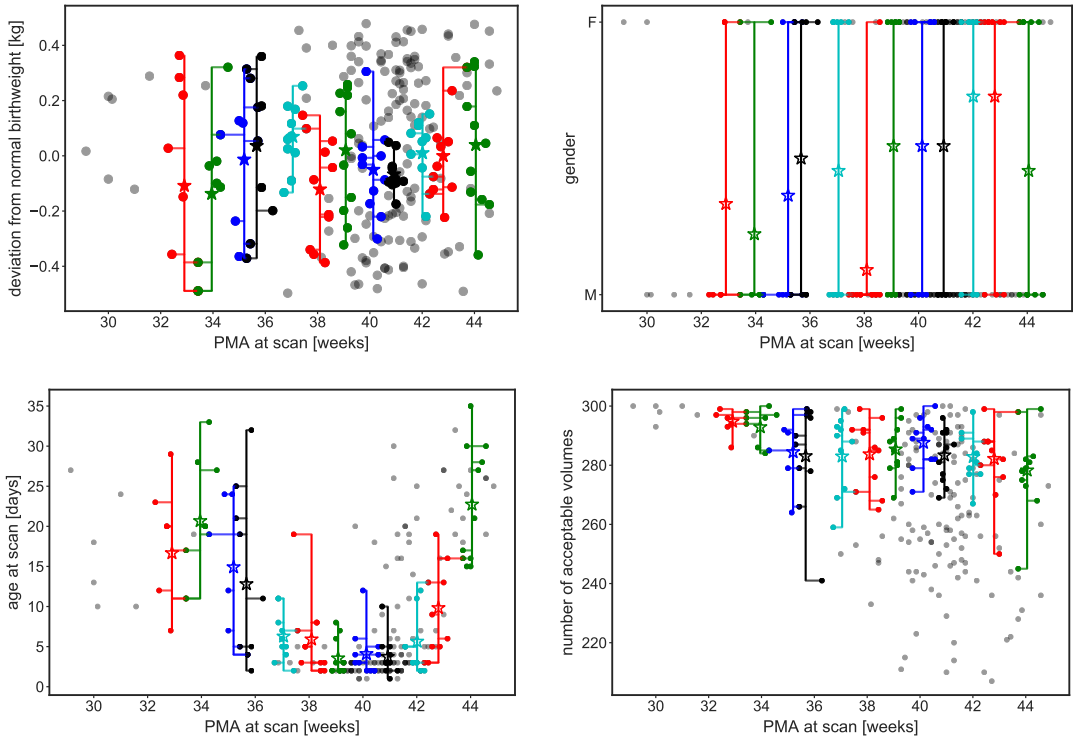


Figure 9.13.: Demographics of the cohort. Lines link subjects that were grouped to build the weekly templates. Asterisks indicate the average for each template. Subjects that are part of the dHCP but were not used for this work are shown as grey dots.

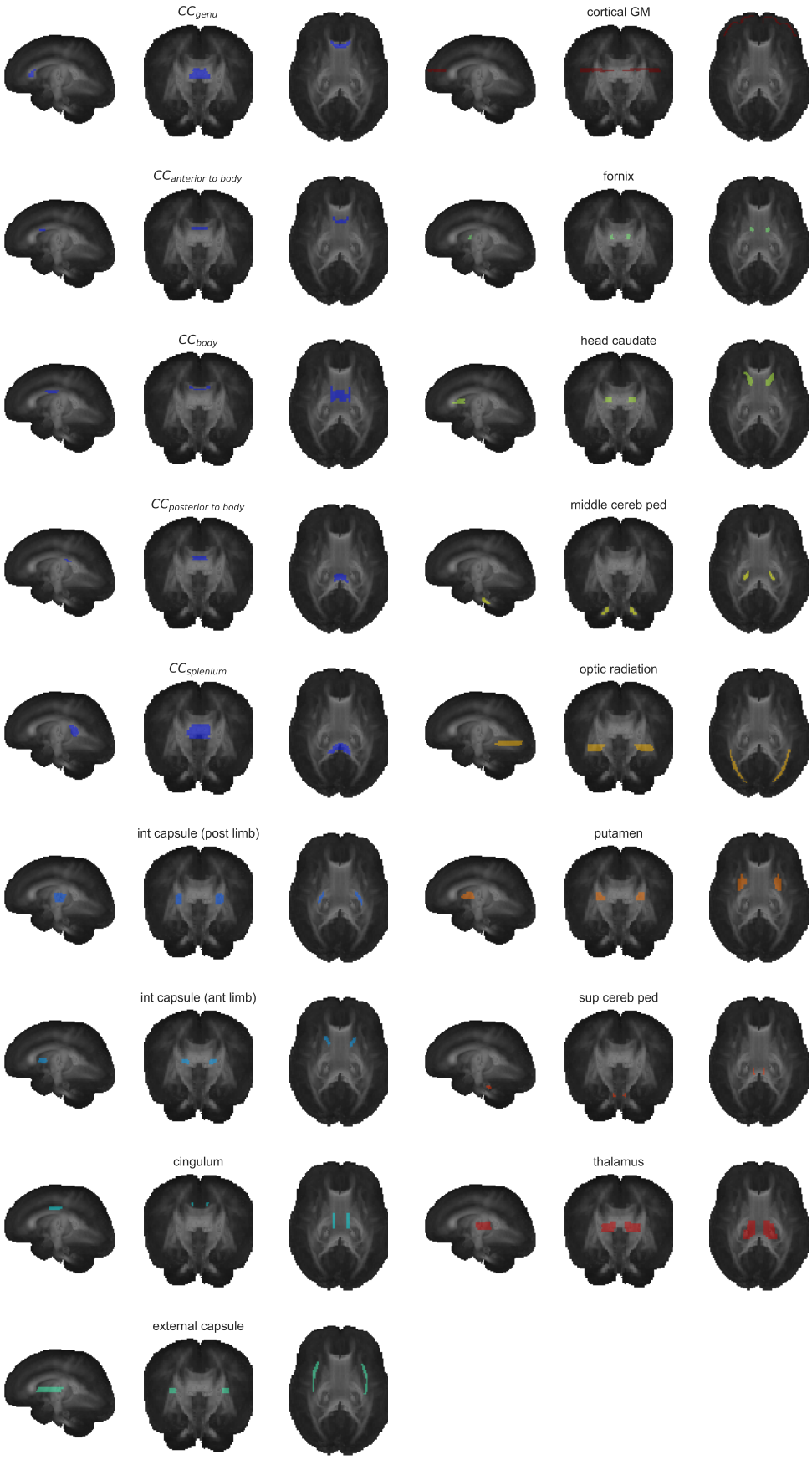


Figure 9.14.: Maximum intensity projections of regions of interest overlaid onto a maximum intensity projection of the age-average FA image.

Chapter 10

Conclusion

Advances in High Angular Resolution Diffusion Imaging (HARDI) facilitates studying neonatal brain development with higher order diffusion models [Hutter et al., 2017; Tournier et al., 2015b]. This technique has the potential to non-invasively probe the cellular make-up and to facilitate the study of normal and abnormal brain maturation in-vivo.

However, foetal and neonatal subjects tend to move during the time required for the acquisition of high-quality HARDI data. Current post-processing tools are not capable of robustly detecting and removing all motion corrupted data. Therefore, in practice, severely artefacted data are manually removed to avoid degrading downstream analysis. This process is labour intensive and subject to rater-variability. To allow large-scale analysis of HARDI data acquired as part of the Developing Human Connectome Project (dHCP), we developed a supervised classification algorithm that accurately distinguishes between motion corrupted and usable data. Currently, the safest approach is to remove corrupted data entirely but ideally, artefact detection should be incorporated in the processing pipeline, automatically judging on the slice or voxel-level, whether data is usable.

To analyse the processed neonatal data, it is necessary to decide on which data representation or microstructure model is best-suited for a specific application. The diffusion tensor representation and its derived quantities have been the method of choice for a large number of studies, despite the ambiguities of interpreting these measures in the presence of multiple fibre populations in a voxel [Wheeler Kingshott, Cercignani, 2009]. In chapter 7, we simulate diffusion in parallel fibres showing that the changes in diffusion tensor quantities during myelination or demyelination are highly dependent on the composition of the tissue, also rendering an interpretation of diffusion tensor quantities in the absence of fibre crossings questionable.

HARDI allows fitting microstructure models that have a higher degrees of freedom than the diffusion tensor model, which allows extracting more informative parameters from the data and potentially facilitates characterising tissue properties with higher specificity. Currently, there exist a number of approaches for deriving tissue characteristics from the data but they are riddled with fundamental flaws with respect to specificity and interpretability [Novikov, Kiselev, Jespersen, 2018]. The brain in the foetal and neonatal period undergoes rapid changes in its shape, size, composition and structure; amplifying

the need for a robust and descriptive tissue model.

We chose to use the constrained spherical deconvolution (CSD) technique and a data-driven approach to extract features ("response functions") that characterise changes in the developing brain. Studies using CSD on neonatal populations have shown the ability to resolve white matter fibre-specific differences between infants born at term and those born prematurely [Pecheva et al., 2017; Blesa et al., 2017; Pannek et al., 2018]. This voxel or fibre-based statistical analysis requires the accurate alignment of subjects to a common space. We extended an existing registration framework to multiple orientation-resolved channels, improving spatial correspondence and angular resolution compared to white matter-based registration.

In adults, it is possible to separate the HARDI signal into cerebrospinal fluid (CSF), white matter, and cortical grey matter components [Jeurissen et al., 2014]. However, in neonates, the overall high water content in the brain, the high anisotropy in cortical grey matter and the heterogeneous developmental state of the brain make this separation into tissue components ambiguous. Therefore, we opted to decompose the signal using the signal fingerprints of free water and white matter. Explicitly taking the high free water content in the developing brain into account, we create a group template of healthy subjects imaged at term that resolves these two characteristics of brain tissue with an unprecedented angular and spatial detail.

A model that captures brain maturation requires features that are sensitive to temporal processes. Investigations into longitudinal signal characteristics reveals that the maturation of white matter can be modelled as an isotropic free-water component and two anisotropic components. There is no guarantee that these patterns are specific to cellular maturation processes but they might be useful in characterising changes in tissue properties occurring during development and maturation. Using this three component decomposition and the CSD framework, we built a high quality atlas of microstructural tissue properties that explicitly represent changing free water content and tissue maturation. This atlas can be directly used to analyse fibre-resolved maturation patterns of the dHCP cohort.

The artefact removal, multi-channel registration and longitudinal modelling of tissue response functions provides a framework for processing and analysing foetal and neonatal diffusion data to study normal and pathological brain maturation at the fibre population-level. Yet, being a data-driven approach, the longitudinal modelling is limited by the age-range studied. An extension to in-utero imaging and HARDI data from older infants could potentially provide a representation that better captures more fundamental tissue properties, such as neuronal maturation and myelination. This could bring us closer to developing a higher order model or signal representation that is specific to cellular processes. Jointly modelling fetal, neonatal and infant data could potentially facilitate extending the time-resolved tissue decomposition to additional contrasts and possibly improve specificity to cortical and deep grey matter.

Future directions of research will aim at applying the decomposition to the full cohort of the dHCP and linking this data to psychological outcome measures. Linking and harmonising HARDI data or HARDI-derived tissue property maps from a large and diverse set of subjects, acquired using different scanner protocols, remains an open research

challenge that could open new pathways for understanding normal and abnormal tissue maturation trajectories and how they relate to outcome.

Bibliography

Abdi, Hervé (2010). "Partial least squares regression and projection on latent structure regression (PLS Regression)". *WIREs Comp Stat* 2.1, pp. 97–106 (page 180).

Adams, Niall M, Hand, David J (1999). "Comparing classifiers when the misallocation costs are uncertain". *Pattern Recognition* 32.7, pp. 1139–1147 (page 79).

Adams, Niall M., Hand, David J. (2000). "Improving the practice of classifier performance assessment". *Neural computation* 12.2, pp. 305–311 (page 79).

Akaike, Hirotogu (1998). "Information theory and an extension of the maximum likelihood principle". *Selected Papers of Hirotugu Akaike*. Springer, pp. 199–213 (page 66).

Akazawa, Kentaro et al. (2016). "Probabilistic maps of the white matter tracts with known associated functions on the neonatal brain atlas: Application to evaluate longitudinal developmental trajectories in term-born and preterm-born infants." *Neuroimage* 128, pp. 167–79. DOI: 10.1016/j.neuroimage.2015.12.026 (page 187).

Akhondi-Asl, Alireza et al. (2015). "Fast myelin water fraction estimation using 2D multislice CPMG". en. *Magnetic Resonance in Medicine* (page 175).

Albert, Monika et al. (2007). "Extensive cortical remyelination in patients with chronic multiple sclerosis". *Brain Pathology* 17.2, pp. 129–138. DOI: 10.1111/j.1750-3639.2006.00043.x (page 169).

Alexander, Daniel C (2008). "A general framework for experiment design in diffusion MRI and its application in measuring direct tissue-microstructure features". *Magnetic Resonance in Medicine* 60.2, pp. 439–448 (page 55).

Alexander, DC, Barker, GJ, Arridge, SR (2002). "Detection and modeling of non-Gaussian apparent diffusion coefficient profiles in human brain data". *Magnetic Resonance in Medicine* 48.2, pp. 331–340 (page 55).

Alexander, DC, Seunarine, KK (2010). "Mathematics of Crossing Fibres". *Diffus. MRI, Theory, Methods Appl. (editor Jones D.K.)* Oxford University Press, USA, p. 456 (page 55).

Amiry-Moghaddam, Mahmood, Ottersen, Ole P (2003). "The molecular basis of water transport in the brain". *Nature Reviews Neuroscience* 4.12, p. 991 (page 52).

Amjad, R. A., Geiger, B. C. (2018). "How (Not) To Train Your Neural Network Using the Information Bottleneck Principle". *ArXiv e-prints*. arXiv: 1802.09766 [cs.LG] (page 73).

Andersen, Susan L (2003). "Trajectories of brain development: point of vulnerability or window of opportunity?" *Neuroscience & Biobehavioral Reviews* 27.1, pp. 3–18 (page 17).

Anderson, Adam W (2005). "Measurement of fiber orientation distributions using high angular resolution diffusion imaging". *Magnetic resonance in medicine* 54.5, pp. 1194–1206 (page 60).

Anderson, Adam W, Gore, John C (1994). "Analysis and correction of motion artifacts in diffusion weighted imaging". *Magnetic resonance in medicine* 32.3, pp. 379–387 (page 47).

Anderson, Adam W et al. (1996). "Effects of osmotically driven cell volume changes on diffusion-weighted imaging of the rat optic nerve". *Magnetic Resonance in Medicine* 35.2, pp. 162–167 (page 170).

Anderson, Philip W et al. (1972). "More is different". *Science* 177.4047, pp. 393–396 (page 67).

Anderson, Stewart A et al. (2001). "Distinct cortical migrations from the medial and lateral ganglionic eminences". *Development* 128.3, pp. 353–363 (page 24).

Andersson, Jesper L R, Sotiropoulos, Stamatios N. (2015a). "Non-parametric representation and prediction of single- and multi-shell diffusion-weighted MRI data using Gaussian processes". *Neuroimage* 122, pp. 166–176. DOI: 10.1016/j.neuroimage.2015.07.067 (page 99).

Andersson, Jesper LR, Jenkinson, Mark, Smith, Stephen, et al. (2007). "Non-linear registration, aka Spatial normalisation FMRIB technical report TR07JA2". *FMRIB Analysis Group of the University of Oxford* 2, pp. 1–21 (page 188).

Andersson, Jesper L.R., Skare, Stefan, Ashburner, John (2003). "How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging". *Neuroimage* 20.2, pp. 870–888. DOI: 10.1016/S1053-8119(03)00336-7 (pages 46, 217).

Andersson, Jesper L.R., Sotiropoulos, Stamatios N. (2015b). "An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging". *Neuroimage* 125, pp. 1063–1078. DOI: 10.1016/j.neuroimage.2015.10.019 (pages 47, 99, 217).

Andersson, Jesper LR et al. (2016). "Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images". *Neuroimage* 141, pp. 556–572 (pages 99, 204).

Andrews, Trevor J, Osborne, Michael T, Does, Mark D (2006). "Diffusion of myelin water". *Magnetic resonance in medicine* 56.2, pp. 381–385 (pages 49, 175).

Andrychowicz, Marcin et al. (2016). "Learning to learn by gradient descent by gradient descent". *Advances in Neural Information Processing Systems*, pp. 3981–3989 (pages 77, 152).

Arlot, Sylvain, Celisse, Alain, et al. (2010). "A survey of cross-validation procedures for model selection". *Statistics surveys* 4, pp. 40–79 (page 93).

Arnold, Vladimir I, Khesin, Boris A (1992). "Topological methods in hydrodynamics". *Annual review of fluid mechanics* 24.1, pp. 145–166 (page 190).

Arsigny, Vincent et al. (2009). "A fast and log-euclidean polyaffine framework for locally linear registration". *Journal of Mathematical Imaging and Vision* 33.2, pp. 222–238 (page 194).

Ascoli, Giorgio A, Donohue, Duncan E, Halavi, Maryam (2007). "NeuroMorpho. Org: a central resource for neuronal morphologies". *Journal of Neuroscience* 27.35, pp. 9247–9251 (page 18).

Ashtari, Manzar et al. (2007). "White matter development during late adolescence in healthy males: a cross-sectional diffusion tensor imaging study." *Neuroimage* 35.2, pp. 501–10. DOI: 10.1016/j.neuroimage.2006.10.047 (page 170).

Assaf, Yaniv, Basser, Peter J (2005). "Composite hindered and restricted model of diffusion (CHARMED) MR imaging of the human brain". *Neuroimage* 27.1, pp. 48–58 (page 55).

Assaf, Yaniv et al. (2008). "AxCaliber: a method for measuring axon diameter distribution from diffusion MRI". *Magnetic resonance in medicine* 59.6, pp. 1347–1354 (pages 57, 174).

Aung, Wint Yan, Mar, Soe, Benzinger, Tammie LS (2013). "Diffusion tensor MRI as a biomarker in axonal and myelin damage". *Imaging in medicine* 5.5, p. 427 (page 169).

Avants, Brian B et al. (2008). "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain". *Medical image analysis* 12.1, pp. 26–41 (pages 190, 191).

Avants, Brian B et al. (2010). "The optimal template effect in hippocampus studies of diseased populations". *Neuroimage* 49.3, pp. 2457–2466 (pages 192, 193).

Avants, Brian B et al. (2015). "The pediatric template of brain perfusion Background & Summary". DOI: 10.1038/sdata.2015.3 (page 187).

Ba, L. J., Caruana, R. (2013). "Do Deep Nets Really Need to be Deep?" *ArXiv e-prints*. arXiv: 1312.6184 [cs.LG] (page 71).

Back, Stephen a et al. (2002). "Arrested oligodendrocyte lineage progression during human cerebral white matter development: dissociation between the timing of progenitor differentiation and myelinogenesis." *Journal of Neuropathology and Experimental Neurology* 61.2, pp. 197–211. DOI: 10.1093/jnen/61.2.197 (page 230).

Back, Stephen A et al. (2001). "Late oligodendrocyte progenitors coincide with the developmental window of vulnerability for human perinatal white matter injury". *Journal of Neuroscience* 21.4, pp. 1302–1312 (page 33).

Baehrens, David et al. (2010). "How to explain individual classification decisions". *Journal of Machine Learning Research* 11.Jun, pp. 1803–1831 (page 156).

Bakiri, Yamina et al. (2011). "Morphological and electrical properties of oligodendrocytes in the white matter of the corpus callosum and cerebellum". *The Journal of physiology* 589.3, pp. 559–573 (page 31).

Balestrini, M R et al. (1991). *Infantile hereditary neuropathy with hypomyelination: report of two siblings with different expressivity* (page 169).

Banwell, B et al. (2009). "Incidence of acquired demyelination of the CNS in Canadian children". *Neurology* 72.3, pp. 232–239. DOI: 10.1212/01.wnl.0000339482.84392.bd (page 169).

Barkovich, A J (2000). "Concepts of myelin and myelination in neuroradiology." *AJNR. Am. J. Neuroradiol.* 21.6, pp. 1099–109 (pages 34, 35, 49).

Barkovich, A. J. et al. (1988). "Normal maturation of the neonatal and infant brain: MR imaging at 1.5 T". *Radiology* 166.1 I, pp. 173–180 (pages 34, 35, 230).

Barres, Ben A, Raff, Martin C (1999). "Axonal control of oligodendrocyte development". *The Journal of cell biology* 147.6, pp. 1123–1128 (page 33).

Barth, Markus et al. (2016). "Simultaneous multislice (SMS) imaging techniques". *Magnetic resonance in medicine* 75.1, pp. 63–81 (page 46).

Bartheld, Christopher S, Bahney, Jami, Herculano-Houzel, Suzana (2016). "The search for true numbers of neurons and glial cells in the human brain: a review of 150 years of cell counting". *Journal of Comparative Neurology* 524.18, pp. 3865–3895 (page 15).

Basser, Peter J (2002). "Relationships between diffusion tensor and q-space MRI". *Magnetic resonance in medicine* 47.2, pp. 392–397 (page 42).

Basser, Peter J, Mattiello, James, LeBihan, Denis (1994). "Estimation of the Effective Self-Diffusion Tensor from the NMR Spin Echo". *Journal of Magnetic Resonance. Series B* 103.3, pp. 247–254 (pages 53, 54).

Basser, Peter J et al. (2000). "In vivo fiber tractography using DT-MRI data". *Magnetic resonance in medicine* 44.4, pp. 625–632 (page 56).

Baumann, Nicole, Pham-Dinh, Danielle (2001). "Biology of oligodendrocyte and myelin in the mammalian central nervous system". *Physiological reviews* 81.2, pp. 871–927 (pages 31, 33).

Bava, Sunita et al. (2010). "Longitudinal characterization of white matter maturation during adolescence." *Brain Research* 1327, pp. 38–46. DOI: 10.1016/j.brainres.2010.02.066 (page 170).

Bayer, Sh A et al. (1993). "Timetables of neurogenesis in the human brain based on experimentally determined patterns in the rat." *Neurotoxicology* 14.1, p. 83 (page 28).

Beaulieu, Christian (2002). "The basis of anisotropic water diffusion in the nervous system–a technical review". *NMR in Biomedicine* 15.7-8, pp. 435–455 (page 49).

Beaulieu, Christian, Allen, Peter S. (1994). "Determinants of anisotropic water diffusion in nerves." *Magnetic Resonance in Medicine* 31.4, pp. 394–400. DOI: 10.1002/mrm.1910310408 (pages 49, 170).

Beg, Mirza Faisal, Khan, Ali (2006). "Computing an average anatomical atlas using LDDMM and geodesic shooting". *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*. IEEE, pp. 1116–1119 (page 193).

Behrens, TEJ, Berg, HJ, Jbabdi, S. (2007). "Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?" *Neuroimage* 34.1, pp. 144–155. DOI: 10.1016/j.neuroimage.2006.09.018 (pages 52, 171).

Bejnordi, Babak Ehteshami et al. (2017). "Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer". *JAMA* 318.22, pp. 2199–2210 (page 64).

Beleites, Claudia et al. (2005). "Variance reduction in estimating classification error using sparse datasets". *Chemometrics and intelligent laboratory systems* 79.1-2, pp. 91–100 (page 94).

Bengio, Yoshua (2011). "Deep Learning of Representations for Unsupervised and Transfer Learning". *JMLR Work. Conf. Proc.* 7, pp. 1–20. DOI: 10.1109/IJCNN.2011.6033302. arXiv: 1606.09549 (pages 63, 67, 71, 77).

– (2012). "Practical recommendations for gradient-based training of deep architectures". *Neural networks: Tricks of the trade*. Springer, pp. 437–478 (page 72).

Bhatia, Kanwal K et al. (2004). "Consistent groupwise non-rigid registration for atlas construction". *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*. IEEE, pp. 908–911 (page 192).

Bignami, Amico, Hosley, Mark, Dahl, Doris (1993). "Hyaluronic acid and hyaluronic acid-binding proteins in brain extracellular matrix". *Anatomy and embryology* 188.5, pp. 419–433 (page 30).

Björk, Marcus et al. (2016). "A multicomponent T2 relaxometry algorithm for myelin water imaging of the brain". en. *Magnetic Resonance in Medicine* 75.1, pp. 390–402 (page 175).

Blaimer, Martin et al. (2013). "Multiband phase-constrained parallel MRI". *Magnetic resonance in medicine* 69.4, pp. 974–980 (page 45).

Blaurock, Allen E (1981). "The spaces between membrane bilayers within PNS myelin as characterized by X-ray diffraction". *Brain research* 210.1-2, pp. 383–387 (page 33).

Blesa, Manuel et al. (2017). "Fixel-based morphometry detects alterations in specific fibres in association with preterm birth: a proof-of-concept study". *ISMRM 25th Annu. Meet. Exhib.* (Page 237).

Bloch, Felix (1946). "Nuclear induction". *Physical review* 70.7-8, p. 460 (pages 39, 43).

Bloy, Luke, Verma, Ragini (2010). "Demons registration of high angular resolution diffusion images". *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*. IEEE, pp. 1013–1016 (page 191).

Blum, Avrim, Rivest, Ronald L (1989). "Training a 3-node neural network is NP-complete". *Advances in neural information processing systems*, pp. 494–501 (page 72).

Bondareff, William, Pysh, Joseph J (1968). "Distribution of the extracellular space during postnatal maturation of rat cerebral cortex". *The Anatomical Record* 160.4, pp. 773–780 (page 30).

Borgnia, Mario et al. (1999). "Cellular and molecular biology of the aquaporin water channels". *Annual review of biochemistry* 68.1, pp. 425–458 (page 53).

Bosman, Fred T, Stamenkovic, Ivan (2003). "Functional structure and composition of the extracellular matrix". *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 200.4, pp. 423–428 (page 30).

Boughorbel, Sabri, Jarray, Fethi, El-Anbari, Mohammed (2017). "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". *PloS one* 12.6, e0177678 (page 80).

Bouix, Sylvain, Rathi, Yogesh, Sabuncu, Mert (2010). "Building an average population HARDI atlas". *MICCAI Workshop on Computational Diffusion MRI*, pp. 84–91 (page 187).

Bouwmans, T. et al. (2016). "On the Role and the Importance of Features for Background Modeling and Foreground Detection". *ArXiv e-prints*. arXiv: 1611.09099 [cs.CV] (page 100).

Boyd, Kendrick, Eng, Kevin H, Page, C David (2013). "Area under the precision-recall curve: Point estimates and confidence intervals". *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 451–466 (pages 82, 95).

Bradley, Andrew P (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". *Pattern recognition* 30.7, pp. 1145–1159 (page 82).

Braga-Neto, Ulisses M, Dougherty, Edward R (2004). "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics* 20.3, pp. 374–380 (page 93).

Branson, Helen M. (2013). "Normal Myelination. A Practical Pictorial Review". *Neuroimaging Clinics of North America* 23.2, pp. 183–195. DOI: 10.1016/j.nic.2012.12.001 (page 225).

Bressloff, Paul C (2014). *Stochastic processes in cell biology*. Vol. 41. Springer (page 38).

Brockstedt, Sara et al. (1999). "Triggering in quantitative diffusion imaging with single-shot EPI". *Acta Radiologica* 40.3, pp. 263–269 (page 48).

Brody, B a et al. (1987). "Sequence of Central Nervous System Myelination in Human Infancy. I. An Autopsy Study of Myelination". *Journal of Neuropathology and Experimental Neurology* 46.3, pp. 283–301. DOI: 10.1097/00005072-198705000-00005 (pages 28, 35, 187, 230).

Brown, Robert (1828). "XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies". *The Philosophical Magazine* 4.21, pp. 161–173 (page 37).

Budde, Matthew D, Annese, Jacopo (2013). "Quantification of anisotropy and fiber orientation in human brain histological sections". en. *Front. Integr. Neurosci.* 7, p. 3 (page 176).

Budde, Matthew D et al. (2009). "Axial diffusivity is the primary correlate of axonal injury in the experimental autoimmune encephalomyelitis spinal cord: a quantitative pixelwise analysis". *Journal of Neuroscience* 29.9, pp. 2805–2813 (page 169).

Bui, Tony et al. (2006). "Microstructural development of human brain assessed in utero by diffusion tensor imaging." *Pediatric Radiology* 36.11, pp. 1133–1140. DOI: 10.1007/s00247-006-0266-3 (page 230).

Buja, Andreas, Stuetzle, Werner, Shen, Yi (2005). "Loss functions for binary class probability estimation and classification: Structure and applications". *Working draft, November* 3 (page 66).

Burkhalter, Andreas, Bernardo, Kerry L, Charles, Vinod (1993). "Development of local circuits in human visual cortex". *Journal of Neuroscience* 13.5, pp. 1916–1931 (pages 17, 30).

Burzynska, A Z et al. (2010). "Age-related differences in white matter microstructure: region-specific patterns of diffusivity". en. *Neuroimage* 49.3, pp. 2104–2112 (page 170).

Bystron, Irina, Blakemore, Colin, Rakic, Pasko (2008). "Development of the human cerebral cortex: Boulder Committee revisited". *Nature Reviews Neuroscience* 9.2, pp. 110–122 (page 22).

Campbell, Jennifer S W et al. (2017). "Promise and pitfalls of g-ratio estimation with MRI". arXiv: 1701.02760 (page 184).

Cardoso, M Jorge et al. (2015). "Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion". *IEEE transactions on medical imaging* 34.9, pp. 1976–1988 (page 192).

Cayre, Myriam, Canoll, Peter, Goldman, James E (2009). "Cell migration in the normal and pathological postnatal mammalian brain". *Progress in neurobiology* 88.1, pp. 41–63 (page 24).

Cercignani, Mara et al. (2017). "Characterizing axonal myelination within the healthy population: a tract-by-tract mapping of effects of age and gender on the fiber g-ratio". *Neurobiology of aging* 49, pp. 109–118 (page 169).

Chang, Eric H et al. (2017). "The role of myelination in measures of white matter integrity: combination of diffusion tensor imaging and two-photon microscopy of CLARITY intact brains". *Neuroimage* 147, pp. 253–261 (page 184).

Chang, Lin-Ching, Jones, Derek K, Pierpaoli, Carlo (2005). "RESTORE: robust estimation of tensors by outlier rejection". *Magnetic resonance in medicine* 53.5, pp. 1088–1095 (page 99).

Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". *Journal of artificial intelligence research* 16, pp. 321–357 (page 76).

Cheng, Guang et al. (2009). "Non-rigid registration of high angular resolution diffusion images represented by gaussian mixture fields". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 190–197 (page 191).

Cheng, Jian et al. (2014). "Non-Negative Spherical Deconvolution (NNSD) for estimation of fiber Orientation Distribution Function in single-/multi-shell diffusion MRI." *Neuroimage* 101, pp. 750–64. DOI: 10.1016/j.neuroimage.2014.07.062 (page 60).

Cheng, Sheung Hun et al. (2001). "Approximating the logarithm of a matrix to specified accuracy". *SIAM Journal on Matrix Analysis and Applications* 22.4, pp. 1112–1125 (page 194).

Cheng, Yen-Wei et al. (2011). "Diffusion tensor imaging with cerebrospinal fluid suppression and signal-to-noise preservation using acquisition combining fluid-attenuated inversion recovery and conventional imaging: comparison of fiber tracking". en. *European Journal of Radiology* 79.1, pp. 113–117 (page 170).

Cheong, J L Y et al. (2009). "Abnormal white matter signal on MR imaging is related to abnormal tissue microstructure." *AJNR. Am. J. Neuroradiol.* 30.3, pp. 623–8. DOI: 10.3174/ajnr.A1399 (pages 169, 170).

Chi, Je G, Dooling, Elizabeth C, Gilles, Floyd H (1977). "Gyral development of the human brain". *Annals of neurology* 1.1, pp. 86–93 (page 15).

Chicco, Davide (2017). "Ten quick tips for machine learning in computational biology". *BioData mining* 10.1, p. 35 (page 80).

Cho, J. et al. (2015). "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?" *ArXiv e-prints*. arXiv: 1511.06348 [cs.LG] (page 102).

Choe, A S et al. (2012). "Validation of diffusion tensor MRI in the central nervous system using light microscopy: quantitative comparison of fiber properties". en. *NMR in Biomedicine* 25.7, pp. 900–908 (page 176).

Chollet, F. (2016). "Xception: Deep Learning with Depthwise Separable Convolutions". *ArXiv e-prints*. arXiv: 1610.02357 [cs.CV] (page 70).

Christiaens, Daan et al. (2015). "Global tractography of multi-shell diffusion-weighted imaging data using a multi-tissue model". *Neuroimage* 123, pp. 89–101 (page 61).

Clancy, B, Darlington, RB, Finlay, BL (2001). "Translating developmental time across mammalian species". *Neuroscience* 105.1, pp. 7–17 (pages 21, 28).

Clark, C A, Barker, G J, Tofts, P S (1999). "An in vivo evaluation of the effects of local magnetic susceptibility-induced gradients on water diffusion measurements in human brain". en. *Journal of Magnetic Resonance* 141.1, pp. 52–61 (page 183).

Codling, Edward A, Plank, Michael J, Benhamou, Simon (2008). "Random walk models in biology." *Journal of the Royal Society, Interface* 5.25, pp. 813–34. DOI: 10.1098/rsif.2008.0014 (page 172).

Coet, Timothy, Suzuki, Kunihiko, Popko, Brian (1998). "New perspectives on the function of myelin galactolipids". *Trends in neurosciences* 21.3, pp. 126–130 (page 34).

Collell, G., Prelec, D., Patil, K. (2016). "Reviving Threshold-Moving: a Simple Plug-in Bagging Ensemble for Binary and Multiclass Imbalanced Data". *ArXiv e-prints*. arXiv: 1606.08698 [cs.LG] (page 76).

Commowick, Olivier, Malandain, Grégoire (2006). "Evaluation of atlas construction strategies in the context of radiotherapy planning". *Proceedings of the SA2PM Workshop (From Statistical Atlases to Personalized Models)*, pp. 1–4 (page 192).

Condon, B et al. (1987). "MR relaxation times of cerebrospinal fluid." *Journal of computer assisted tomography* 11.2, pp. 203–207 (page 48).

Cook, P A A et al. (2006). "Camino: open-source diffusion-MRI reconstruction and processing". *14th Scientific Meeting of the International Society for Magnetic Resonance in Medicine*, p. 2759 (page 172).

Cootes, Timothy F et al. (2004). "Groupwise diffeomorphic non-rigid registration for automatic model building". *European conference on computer vision*. Springer, pp. 316–327 (page 190).

Corbin, Joshua G, Nery, Susana, Fishell, Gord (2001). "Telencephalic cells take a tangent: non-radial migration in the mammalian forebrain". *Nature neuroscience* 4, p. 1177 (page 24).

Courchesne, Eric et al. (2007). "Mapping early brain development in autism". *Neuron* 56.2, pp. 399–413 (page 24).

Cragg, B (1979). "Brain extracellular space fixed for electron microscopy". *Neuroscience letters* 15.2-3, pp. 301–306 (page 30).

Crum, William R, Hartkens, Thomas, Hill, DLG (2004). "Non-rigid image registration: theory and practice". *The British journal of radiology* 77.suppl_2, S140–S153 (page 188).

Cybenko, George (1989). "Approximation by superpositions of a sigmoidal function". *Mathematics of Control, Signals, and Systems (MCSS)* 2.4, pp. 303–314 (page 71).

Daducci, Alessandro et al. (2015). "COMMIT: Convex Optimization Modeling for Microstructure Informed Tractography". *IEEE Transactions on Medical Imaging* 34.1, pp. 246–257. DOI: 10.1109/TMI.2014.2352414 (page 232).

De Craene, Mathieu et al. (2004). "Multi-subject registration for unbiased statistical atlas construction". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 655–662 (page 192).

De Santis, Silvia et al. (2016). "Resolving relaxometry and diffusion properties within the same voxel in the presence of crossing fibres by combining inversion recovery and diffusion-weighted acquisitions". *Magnetic Resonance in Medicine* 75.1, pp. 372–380. DOI: 10.1002/mrm.25644 (page 232).

Dean, Douglas C et al. (2015). "Characterizing longitudinal white matter development during early childhood". *Brain Structure and Function* 220.4, pp. 1921–1933 (page 35).

Dean III, Douglas C et al. (2016). "Mapping an index of the myelin g-ratio in infants using magnetic resonance imaging". *Neuroimage* 132, pp. 225–237 (pages 169, 184).

DeAzevedo, Leonardo C, Hedin-Pereira, Cecilia, Lent, Roberto (1997). "Callosal neurons in the cingulate cortical plate and subplate of human fetuses". *Journal of Comparative Neurology* 386.1, pp. 60–70 (page 28).

Debanne, Dominique (2004). "Information processing in the axon". *Nature Reviews Neuroscience* 5.4, p. 304 (page 18).

Deese, ALAN J et al. (1982). "Proton NMR T1, T2, and T1 rho relaxation studies of native and reconstituted sarcoplasmic reticulum and phospholipid vesicles". *Biophysical journal* 37.1, pp. 207–216 (page 49).

Dehaene-Lambertz, G., Spelke, E.S. (2015). "The Infancy of the Human Brain". *Neuron* 88.1, pp. 93–109. DOI: 10.1016/j.neuron.2015.09.026 (pages 16, 18, 35).

Dell'Acqua, Flavio et al. (2010). "A modified damped Richardson–Lucy algorithm to reduce isotropic background effects in spherical deconvolution". *Neuroimage* 49.2, pp. 1446–1458 (page 60).

Demšar, Janez (2006). "Statistical comparisons of classifiers over multiple data sets". *Journal of Machine learning research* 7.Jan, pp. 1–30 (pages 80, 93–95).

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pp. 248–255 (page 102).

Denninger, Andrew R et al. (2014). "Neutron scattering from myelin revisited: bilayer asymmetry and water-exchange kinetics". *Acta Crystallographica Section D: Biological Crystallography* 70.12, pp. 3198–3211 (pages 33, 50).

Descoteaux, Maxime et al. (2006). "Apparent diffusion coefficients from high angular resolution diffusion imaging: Estimation and applications". *Magnetic Resonance in Medicine* 56.2, pp. 395–410 (pages 55, 59).

Dhollander, Thijs, Raffelt, David, Connelly, Alan (2016). "Unsupervised 3-tissue response function estimation from single-shell or multi-shell diffusion MR data without a co-registered T1 image". *Proc ISMRM Workshop on Breaking the Barriers of Diffusion MRI*. Vol. 5 (pages 61, 196, 219).

Dietrich, Rosalind B et al. (1988). "MR evaluation of early myelination patterns in normal and developmentally delayed infants". *American Journal of Roentgenology* 150.4, pp. 889–896 (page 35).

Dietterich, Thomas G (1998). "Approximate statistical tests for comparing supervised classification learning algorithms". *Neural computation* 10.7, pp. 1895–1923 (page 94).

Dimou, Leda et al. (2008). "Progeny of Olig2-expressing progenitors in the gray and white matter of the adult mouse cerebral cortex". *Journal of Neuroscience* 28.41, pp. 10434–10442 (page 19).

Dinh, Laurent et al. (2017). "Sharp minima can generalize for deep nets". *arXiv preprint arXiv:1703.04933* (pages 72, 120).

Dittrich, Eva et al. (2014). "A spatio-temporal latent atlas for semi-supervised learning of fetal brain segmentations and morphological age estimation". *Medical Image Analysis* 18, pp. 9–21. DOI: 10.1016/j.media.2013.08.004 (page 187).

Dobbing, John, Sands, Jean (1973). "Quantitative growth and development of human brain". *Archives of disease in childhood* 48.10, pp. 757–767 (pages 14, 187, 225).

Donahue, Chad J et al. (2016). "Using diffusion tractography to predict cortical connection strength and distance: a quantitative comparison with tracers in the monkey". *Journal of Neuroscience* 36.25, pp. 6758–6770 (page 56).

Dortch, Richard D et al. (2013). "Characterizing inter-compartmental water exchange in myelinated tissue using relaxation exchange spectroscopy". *Magnetic resonance in medicine* 70.5, pp. 1450–1459 (page 52).

Dougherty, Edward R et al. (2010). "Performance of error estimators for classification". *Current Bioinformatics* 5.1, pp. 53–67 (pages 92, 93).

Drobnjak, Ivana et al. (2006). "Development of a functional magnetic resonance imaging simulator for modeling realistic rigid-body motion artifacts". *Magnetic Resonance in Medicine* 56.2, pp. 364–380 (page 100).

Drummond, Chris, Japkowicz, Nathalie (2010). "Warning: statistical benchmarking is addictive. Kicking the habit in machine learning". *Journal of Experimental & Theoretical Artificial Intelligence* 22.1, pp. 67–80 (pages 79, 95).

Dubois, Jessica et al. (2014). "The early development of brain white matter: a review of imaging studies in fetuses, newborns and infants". *Neuroscience* 276, pp. 48–71 (pages 17, 24, 28).

Dudink, Jeroen et al. (2007). "Fractional anisotropy in white matter tracts of very-low-birth-weight infants". *Pediatric radiology* 37.12, pp. 1216–1223 (page 99).

Ebisu, T et al. (1993). "Discrimination between different types of white matter edema with diffusion-weighted MR imaging". *Journal of Magnetic Resonance Imaging* 3.6, pp. 863–868 (page 170).

Edelstein, William A et al. (1980). "Spin warp NMR imaging and applications to human whole-body imaging". *Physics in medicine & biology* 25.4, p. 751 (page 40).

Edgar, Julia M, Griffiths, Ian R (2009). "White matter structure: a microscopist's view". *Diffusion Mri*. Elsevier, pp. 74–103 (page 49).

Efron, Bradley (1983). "Estimating the error rate of a prediction rule: improvement on cross-validation". *Journal of the American statistical association* 78.382, pp. 316–331 (page 93).

– (2004). "The estimation of prediction error: covariance penalties and cross-validation". *Journal of the American Statistical Association* 99.467, pp. 619–632 (page 93).

Efron, Bradley, Tibshirani, Robert (1986). "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy". *Statistical science*, pp. 54–75 (pages 93, 105).

– (1997). "Improvements on cross-validation: the 632+ bootstrap method". *Journal of the American Statistical Association* 92.438, pp. 548–560 (pages 79, 93).

Eickenberg, Michael et al. (2017). "Seeing it all: Convolutional network layers map the function of the human visual system". *Neuroimage* 152, pp. 184–194 (page 67).

Einstein, Albert (1905). "On the movement of small particles suspended in a stationary liquid demanded by the molecular-kinetic theory of heat". *Annales de Physique* 17.8, pp. 549–560. DOI: 10.1002/andp.19053220806 (page 37).

Elsken, T., Metzen, J.-H., Hutter, F. (2017). "Simple And Efficient Architecture Search for Convolutional Neural Networks". *ArXiv e-prints*. arXiv: 1711.04528 [stat.ML] (pages 151, 152).

Engle, William A (2004). "Age terminology during the perinatal period." *Pediatrics* 114.5, pp. 1362–1364 (page 15).

Esteva, Andre et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". *Nature* 542.7639, pp. 115–118 (page 108).

Evans, Alan C et al. (2012). "Brain templates and atlases". *Neuroimage* 62.2, pp. 911–922 (pages 192, 199).

Feldman, Heidi M et al. (2010). "Diffusion Tensor Imaging: A Review for Pediatric Researchers and Clinicians". *Journal of Developmental and Behavioral Pediatrics* 31.4, pp. 346–356. DOI: 10.1097/DBP.0b013e3181dcaa8b.Diffusion (pages 169, 170).

Ferguson, B et al. (1997). "Axonal damage in acute multiple sclerosis lesions". en. *Brain* 120 (Pt 3), pp. 393–399 (page 170).

Ferizi, Uran et al. (2015). "White matter compartment models for in vivo diffusion MRI at 300 mT/m". *Neuroimage* 118, pp. 468–483. DOI: 10.1016/j.neuroimage.2015.06.027 (pages 57, 58).

Ferrante, Enzo, Paragios, Nikos (2017). "Slice-to-volume medical image registration: A survey". *Medical image analysis* 39, pp. 101–123 (page 188).

Ferri, Cesar, Hernández-Orallo, José, Flach, Peter A (2011). "A coherent interpretation of AUC as a measure of aggregated classification performance". *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 657–664 (page 83).

Fick, Adolf (1855). "Ueber diffusion". *Annalen der Physik* 170.1, pp. 59–86 (page 37).

Fields, R Douglas (2014). "Myelin—more than insulation". *Science* 344.6181, pp. 264–266 (page 31).

– (2005). "Myelination: an overlooked mechanism of synaptic plasticity?" *The Neuroscientist* 11.6, pp. 528–531 (page 35).

– (2008). *White matter in learning, cognition and psychiatric disorders*. DOI: 10.1016/j.tins.2008.04.001. arXiv: NIHMS150003 (page 169).

Fieremans, Els et al. (2012). "Diffusion distinguishes between axonal loss and demyelination in brain white matter". *20th Annual Meeting of the International Society for Magnetic Resonance in Medicine. Melbourne, Australia*. cds.ismrm.org, p. 465 (page 171).

Fisher, Elizabeth et al. (2007). "Imaging correlates of axonal swelling in chronic multiple sclerosis brains". en. *Annals of Neurology* 62.3, pp. 219–228 (page 170).

Flechsig (2017). *File:FlechsigSaggital4.jpg — Wikipedia, The Free Encyclopedia*. [Online; accessed 23-February-2018] (page 35).

Flechsig, P.E. (1920). *Anatomie des menschlichen Gehirns und Rückenmarks auf myelogenetischer Grundlage*. Anatomie des menschlichen Gehirns und Ruckenmarks auf myelogenetischer Grundlage v. 1. G. Thieme (page 35).

Forman, George, Scholz, Martin (2010). "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement". *ACM SIGKDD Explorations Newsletter* 12.1, pp. 49–57 (pages 79, 94, 95).

Fox, R J et al. (2011). "Measuring myelin repair and axonal loss with diffusion tensor imaging." *AJNR. Am. J. Neuroradiol.* 32.1, pp. 85–91. DOI: 10.3174/ajnr.A2238 (pages 169–171, 182).

Frank, Lawrence R (2002). "Characterization of anisotropy in high angular resolution diffusion-weighted MRI". *Magnetic Resonance in Medicine* 47.6, pp. 1083–1099 (page 60).

Freund, Tamas F, Buzsáki, Gyorgi (1996). "Interneurons of the hippocampus". *Hippocampus* 6.4, pp. 347–470 (page 18).

Freund, Yoav, Schapire, Robert E (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of computer and system sciences* 55.1, pp. 119–139 (page 144).

Friede, Reinhard L (1972). "Control of myelin formation by axon caliber.(With a model of the control mechanism)". *Journal of Comparative Neurology* 144.2, pp. 233–252 (page 31).

Fukushima, Kunihiko (1979). "Neural network model for a mechanism of pattern recognition unaffected by shift in position-Neocognitron". *IEICE Technical Report, A* 62.10, pp. 658–665 (page 67).

Galar, Mikel et al. (2012). "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches". *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.4, pp. 463–484 (page 76).

Gall, J. C. et al. (1958). "Multiple sclerosis in children; a clinical study of 40 cases with onset in childhood." *Pediatrics* 21.5, pp. 703–9 (page 169).

Geiger, B. C., Feldbauer, C., Kubin, G. (2011). "Information Loss in Static Nonlinearities". *ArXiv e-prints*. arXiv: 1102.4794 [cs.IT] (page 67).

Geng, Xiujuan et al. (2011). "Diffeomorphic image registration of diffusion MRI using spherical harmonics". *IEEE transactions on medical imaging* 30.3, pp. 747–758 (page 191).

Ghashghaei, H Troy, Lai, Cary, Anton, ES (2007). "Neuronal migration in the adult brain: are we there yet?" *Nature Reviews Neuroscience* 8.2, p. 141 (pages 20, 24).

Gibson, Erin M et al. (2014). "Neuronal activity promotes oligodendrogenesis and adaptive myelination in the mammalian brain". *Science* 344.6183, p. 1252304 (page 34).

Gilles, FH, Shankle, W, Dooling, EC (1983). "Myelinated tracts: growth patterns". *The developing human brain*. Elsevier, pp. 117–183 (pages 35, 230).

Girard, Gabriel et al. (2014). "Towards quantitative connectivity analysis: reducing tractography biases". *Neuroimage* 98, pp. 266–278 (page 56).

Glasser, Matthew F et al. (2013). "The minimal preprocessing pipelines for the Human Connectome Project". *Neuroimage* 80, pp. 105–124 (page 196).

Glorot, Xavier, Bengio, Yoshua (2010). "Understanding the difficulty of training deep feedforward neural networks". *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256 (page 118).

Gneiting, Tilmann, Raftery, Adrian E (2007). "Strictly proper scoring rules, prediction, and estimation". *Journal of the American Statistical Association* 102.477, pp. 359–378 (page 81).

Golabchi, Fatemeh N et al. (2010). "Pixel-based comparison of spinal cord MR diffusion anisotropy with axon packing parameters." *Magnetic Resonance in Medicine* 63.6, pp. 1510–9. DOI: 10.1002/mrm.22337 (page 170).

Goldman, L, Albus, James S (1968). "Computation of impulse conduction in myelinated fibers; theoretical basis of the velocity-diameter relation". *Biophysical journal* 8.5, pp. 596–607 (page 183).

Gong, Gaolang et al. (2008). "Mapping anatomical connectivity patterns of human cerebral cortex using in vivo diffusion tensor imaging tractography". *Cerebral cortex* 19.3, pp. 524–536 (page 56).

Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press (pages 63, 64, 66, 67, 71–73, 119).

Goodlett, Casey B et al. (2009). "Group analysis of DTI fiber tract statistics with application to neurodevelopment". *Neuroimage* 45.1, S133–S142 (page 187).

Greitz, Dan et al. (1992). "Pulsatile brain movement and associated hydrodynamics studied by magnetic resonance phase imaging". *Neuroradiology* 34.5, pp. 370–380 (page 48).

Griswold, Mark A et al. (2002). "Generalized autocalibrating partially parallel acquisitions (GRAPPA)". *Magnetic resonance in medicine* 47.6, pp. 1202–1210 (page 45).

Gu, M. et al. (2009). "More really is different". *Physica D Nonlinear Phenomena* 238, pp. 835–839. DOI: 10.1016/j.physd.2008.12.016. arXiv: 0809.0151 [cond-mat.other] (page 67).

Guillery, RW (2005). "Is postnatal neocortical maturation hierarchical?" *Trends in neurosciences* 28.10, pp. 512–517 (page 35).

Guimond, Alexandre, Meunier, Jean, Thirion, Jean-Philippe (2000). "Average brain models: A convergence study". *Computer vision and image understanding* 77.2, pp. 192–210 (page 192).

Gürbüzbalaban, Mert, Ozdaglar, Asu, Parrilo, Pablo (2015). "Why Random Reshuffling Beats Stochastic Gradient Descent". arXiv: 1510.08560 (page 105).

Habas, Piotr A et al. (2009). "A spatio-temporal atlas of the human fetal brain with application to tissue segmentation." *Med. Image Comput. Comput. Assist. Interv.* 12.Pt 1, pp. 289–96 (page 187).

Hahn, Erwin L (1960). "Detection of sea-water motion by nuclear precession". *Journal of geophysical research* 65.2, pp. 776–777 (page 47).

– (1950). "Spin echoes". *Physical review* 80.4, p. 580 (page 41).

Halevy, Alon, Norvig, Peter, Pereira, Fernando (2009). "The unreasonable effectiveness of data". *IEEE Intelligent Systems* 24.2, pp. 8–12 (page 94).

Hall, John E (2015). *Guyton and Hall textbook of medical physiology e-Book*. Elsevier Health Sciences (page 18).

Hall, Matt G, Alexander, Daniel C (2009). "Convergence and parameter choice for Monte-Carlo simulations of diffusion MRI". *IEEE Transactions on Medical Imaging* 28.9, pp. 1354–1364 (pages 171, 172, 174, 175).

Hand, David J (2009). "Measuring classifier performance: a coherent alternative to the area under the ROC curve". *Machine learning* 77.1, pp. 103–123 (pages 79, 81–83).

Hand, David J, Anagnostopoulos, Christoforos (2014). "A better Beta for the H measure of classification performance". *Pattern Recognition Letters* 40, pp. 41–46 (pages 83, 92).

Hand, David J et al. (2006). "Classifier technology and the illusion of progress". *Statistical science* 21.1, pp. 1–14 (pages 95, 132, 152).

Hardy, Rebecca J (1997). "Dorsoventral patterning and oligodendroglial specification in the developing central nervous system". *Journal of neuroscience research* 50.2, pp. 139–145 (page 33).

Hardy, Rebecca J, Friedrich Jr, Victor L (1996). "Progressive remodeling of the oligodendrocyte process arbor during myelinogenesis". *Developmental neuroscience* 18.4, pp. 243–254 (pages 33, 34).

Harkins, Kevin D, Dula, Adrienne N, Does, Mark D (2012). "Effect of intercompartmental water exchange on the apparent myelin water fraction in multiexponential T2 measurements of rat spinal cord". *Magnetic resonance in medicine* 67.3, pp. 793–800 (page 52).

Harrell, Frank E (1998). "Comparison of strategies for validating binary logistic regression models". *[Accessed April 2018]* (page 94).

Harrell, Frank E, Lee, Kerry L, Mark, Daniel B (1996). "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". *Statistics in medicine* 15.4, pp. 361–387 (pages 92, 93).

Haselgrove, John C, Moore, James R (1996). "Correction for distortion of echo-planar images used to calculate the apparent diffusion coefficient". *Magnetic Resonance in Medicine* 36.6, pp. 960–964 (page 47).

Hassabis, Demis et al. (2017). "Neuroscience-inspired artificial intelligence". *Neuron* 95.2, pp. 245–258 (page 63).

Hastie, Trevor, Friedman, Jerome, Tibshirani, Robert (2009a). "Additive Models, Trees, and Related Methods". *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, pp. 193–224. DOI: 10.1007/978-0-387-21606-5_7 (page 144).

– (2009b). "Model Assessment and Selection". *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, pp. 193–224. DOI: 10.1007/978-0-387-21606-5_7 (page 92).

Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome (2009a). "Overview of supervised learning". *The elements of statistical learning*. Springer, pp. 9–41 (page 64).

– (2009b). *The Elements of Statistical Learning*. Springer New York. DOI: 10.1007/978-0-387-84858-7 (page 65).

He, Haibo, Garcia, Edwardo A (2009). "Learning from imbalanced data". *IEEE Transactions on knowledge and data engineering* 21.9, pp. 1263–1284 (pages 74, 76, 79).

He, Haibo et al. (2008). "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". *Proc. Int. Jt. Conf. Neural Networks* 3, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969 (page 76).

He, K. et al. (2015a). "Deep Residual Learning for Image Recognition". *ArXiv e-prints*. arXiv: 1512.03385 [cs.CV] (pages 71, 73).

He, K. et al. (2015b). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". *ArXiv e-prints*. arXiv: 1502.01852 [cs.CV] (page 119).

Healy Jr, Dennis M, Hendriks, Harrie, Kim, Peter T (1998). "Spherical deconvolution". *Journal of Multivariate Analysis* 67.1, pp. 1–22 (page 59).

Heemskerk, AM et al. (2013). "Acquisition guidelines and quality assessment tools for analyzing neonatal diffusion tensor MRI data". *American Journal of Neuroradiology* 34.8, pp. 1496–1505 (page 99).

Helenius, Johanna et al. (2002). "Diffusion-weighted MR imaging in normal human brains in various age groups". en. *AJNR: American Journal of Neuroradiology* 23.2, pp. 194–199 (page 182).

Henkelman, RM, Stanisz, GJ, Graham, SJ (2001). "Magnetization transfer in MRI: a review". *NMR in Biomedicine* 14.2, pp. 57–64 (page 49).

Hennel, Franciszek et al. (2016). "SENSE reconstruction for multiband EPI including slice-dependent N/2 ghost correction". *Magnetic Resonance in Medicine* 76.3, pp. 873–879. DOI: 10.1002/mrm.25915 (pages 115, 217).

Hernandez, Monica, Olmos, Salvador, Pennec, Xavier (2008). "Comparing algorithms for diffeomorphic registration: Stationary LDDMM and Diffeomorphic Demons". *2nd MICCAI workshop on mathematical foundations of computational anatomy*, pp. 24–35 (page 190).

Hevner, Robert F (2000). "Development of connections in the human visual system during fetal mid-gestation: a Dil-tracing study". *Journal of Neuropathology & Experimental Neurology* 59.5, pp. 385–392 (page 16).

Hidalgo-Tobon, SS (2010). "Theory of gradient coil design methods for magnetic resonance imaging". *Concepts in Magnetic Resonance Part A* 36.4, pp. 223–242 (page 47).

Hildebrand, C, Hahn, R (1978). "Relation between myelin sheath thickness and axon size in spinal cord white matter of some vertebrate species". *Journal of the Neurological Sciences* (page 183).

Hinton, Geoffrey E (1986). "Learning distributed representations of concepts". *Proceedings of the eighth annual conference of the cognitive science society*. Vol. 1. Amherst, MA, p. 12 (page 67).

Hochreiter, Sepp, Younger, A Steven, Conwell, Peter R (2001). "Learning to learn using gradient descent". *International Conference on Artificial Neural Networks*. Springer, pp. 87–94 (page 76).

Holden, Mark (2008). "A review of geometric transformations for nonrigid body registration". *IEEE transactions on medical imaging* 27.1, pp. 111–128 (page 190).

Hong, Xiaole, Thomas Dixon, W (1992). "Measuring diffusion in inhomogeneous systems in imaging mode using antisymmetric sensitizing gradients". *Journal of Magnetic Resonance* 99.3, pp. 561–570 (page 183).

Honig, Lawrence S, Herrmann, Kathrin, Shatz, Carla J (1996). "Developmental changes revealed by immunohistochemical markers in human cerebral cortex". *Cerebral Cortex* 6.6, pp. 794–806 (pages 16, 21).

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". *Neural networks* 4.2, pp. 251–257 (page 71).

Hornik, Kurt, Stinchcombe, Maxwell, White, Halbert (1989). "Multilayer feedforward networks are universal approximators". *Neural networks* 2.5, pp. 359–366 (page 71).

Horssen, Jack van et al. (2006). "Extensive extracellular matrix depositions in active multiple sclerosis lesions". *Neurobiology of disease* 24.3, pp. 484–491 (page 30).

Huang, ZJ, Di Cristo, G, Ango, F (2007). "Development of GABA innervation in the cerebral and cerebellar cortices". *Nature Reviews Neuroscience* 8.9, p. 673 (page 19).

Hughes, Emer J et al. (2017a). "A dedicated neonatal brain imaging system". en. *Magnetic Resonance in Medicine* 78.2, pp. 794–804 (pages 115, 203, 217).

Hughes, Emer J et al. (2017b). "The type and prevalence of incidental findings on magnetic resonance imaging of the low risk term born neonatal brain". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 25, p. 4107 (page 217).

Hughes, Ethan G et al. (2013). "Oligodendrocyte progenitors balance growth with self-repulsion to achieve homeostasis in the adult brain". *Nature neuroscience* 16.6, p. 668 (page 19).

Hüppi, Petra S, Dubois, Jessica (2006). "Diffusion tensor imaging of brain development". *Seminars in Fetal and Neonatal Medicine*. Vol. 11. 6. Elsevier, pp. 489–497 (page 10).

Hurley, S A, Mossahebi, P M (2010). "Multicomponent relaxometry (mcDESPOT) in the shaking pup model of dysmyelination". *Proceedings of the 18th Meeting of the International Society for Magnetic Resonance in Medicine* (page 175).

Hutter, Jana et al. (2017). "Time-efficient and flexible design of optimized multishell HARDI diffusion". *Magnetic resonance in medicine* (pages 44, 115, 116, 204, 217, 236).

Iglesias, Juan Eugenio, Sabuncu, Mert R (2015). "Multi-atlas segmentation of biomedical images: a survey". *Medical image analysis* 24.1, pp. 205–219 (page 192).

Innocenti, Giorgio M, Price, David J (2005). "Exuberance in the development of cortical networks". *Nature Reviews Neuroscience* 6.12, p. 955 (page 30).

Ioffe, Sergey, Szegedy, Christian (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach, David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 448–456 (pages 73, 125, 126).

Ivanova, Maria V et al. (2016). "Diffusion-tensor imaging of major white matter tracts and their role in language processing in aphasia". en. *Cortex* (page 182).

Jaderberg, M. et al. (2017). "Population Based Training of Neural Networks". *ArXiv e-prints*. arXiv: 1711.09846 [cs.LG] (page 152).

Jain, Anil K, Dubes, Richard C, Chen, Chaur-Chin (1987). "Bootstrap techniques for error estimation". *IEEE transactions on pattern analysis and machine intelligence* 5, pp. 628–633 (page 93).

Jäkel, Sarah, Dimou, Leda (2017). "Glial cells and their function in the adult brain: a journey through the history of their ablation". *Frontiers in cellular neuroscience* 11, p. 24 (page 19).

Jamain, Adrien, Hand, David J (2008). "Mining supervised classification performance studies: A meta-analytic investigation". *Journal of Classification* 25.1, pp. 87–112 (page 79).

Japkowicz, Nathalie (2000). "The class imbalance problem: Significance and strategies". *Proc. of the Int'l Conf. on Artificial Intelligence* (page 76).

Japkowicz, Nathalie, Shah, Mohak (2011). *Evaluating learning algorithms: a classification perspective*. Cambridge University Press (page 79).

Jastrzębski, S et al. (2017). "Three Factors Influencing Minima in SGD". *ArXiv e-prints*. arXiv: 1711.04623 [cs.LG] (pages 73, 120).

Jelescu, Ileana O, Budde, Matthew D (2017). "Design and validation of diffusion MRI models of white matter". *Frontiers in Physics* 5, p. 61 (pages 57, 58).

Jelescu, Ileana O et al. (2016). "In vivo quantification of demyelination and recovery using compartment-specific diffusion MRI metrics validated by electron microscopy". *Neuroimage* 132, pp. 104–114 (page 184).

Jensen, Jens H et al. (2005). "Diffusional kurtosis imaging: The quantification of non-gaussian water diffusion by means of magnetic resonance imaging". *Magnetic resonance in medicine* 53.6, pp. 1432–1440 (page 55).

Jespersen, Sune N et al. (2007). "Modeling dendrite density from magnetic resonance diffusion measurements". *Neuroimage* 34.4, pp. 1473–1486 (pages 53, 57).

Jeurissen, Ben, Tournier, Jacques-Donald, Sijbers, Jan (2015). "Tissue-type segmentation using non-negative matrix factorization of multi-shell diffusion-weighted MRI images". *23th ISMRM*, p. 0346 (page 61).

Jeurissen, Ben et al. (2017). "Diffusion MRI fiber tractography of the brain". *NMR in Biomedicine* (page 56).

Jeurissen, Ben et al. (2013). "Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging." *Human Brain Mapping* 34.11, pp. 2747–2766 (pages 52, 171).

Jeurissen, Ben et al. (2014). "Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data". *Neuroimage* 103, pp. 411–426. DOI: 10.1016/j.neuroimage.2014.07.061 (pages 55, 61, 196, 216, 217, 237).

Jindal, I., Nokleby, M., Chen, X. (2017). "Learning Deep Networks from Noisy Labels with Dropout Regularization". *ArXiv e-prints*. arXiv: 1705.03419 [cs.CV] (page 73).

Jones, Derek K, Cercignani, Mara (2010). "Twenty-five pitfalls in the analysis of diffusion MRI data". *NMR in Biomedicine* 23.7. Ed. by Jens H Jensen, Joseph A Helpern, pp. 803–820 (page 169).

Jones, Derek K, Knösche, Thomas R, Turner, Robert (2013). "White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI". *Neuroimage* 73, pp. 239–254 (page 169).

Joshi, Sarang et al. (2004). "Unbiased diffeomorphic atlas construction for computational anatomy". *Neuroimage* 23, S151–S160 (pages 192, 193).

Judaš, Miloš et al. (2005). "Structural, immunocytochemical, and MR imaging properties of periventricular crossroads of growing cortical pathways in preterm infants". *American journal of neuroradiology* 26.10, pp. 2671–2684 (pages 30, 206).

Kaden, Enrico et al. (2016). "Multi-compartment microscopic diffusion imaging". *Neuroimage* 139, pp. 346–359 (page 58).

Kalf, Hubert et al. (1995). "On the expansion of a function in terms of spherical harmonics in arbitrary dimensions". *Bulletin of the Belgian Mathematical Society-Simon Stevin* 2.4, pp. 361–380 (page 59).

Kanold, Patrick O, Luhmann, Heiko J (2010). "The subplate and early cortical circuits". *Annual review of neuroscience* 33, pp. 23–48 (pages 24, 28).

Karampinos, Dimitrios C et al. (2008). "High resolution reduced-FOV diffusion tensor imaging of the human pons with multi-shot variable density spiral at 3T". en. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2008, pp. 5761–5764 (page 170).

Karlicek Jr, RF, Lowe, IJ (1980). "A modified pulsed gradient technique for measuring diffusion in the presence of large background gradients". *Journal of Magnetic Resonance (1969)* 37.1, pp. 75–91 (page 43).

Kelly, Christopher et al. (2017). "Transfer learning and convolutional neural net fusion for motion artefact detection". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 25, p. 3523 (pages 11, 100, 120, 149, 156, 204, 217).

Kettemann, H, Ransom, BR (2005). "The concept of neuroglia: a historical perspective". *Neuroglia, 2nd edn. Oxford University Press, Oxford*, pp. 1–18 (page 19).

Khoshgoftaar, Taghi M., Van Hulse, Jason, Napolitano, Amri (2011). "Comparing boosting and bagging techniques with noisy and imbalanced data". *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans* 41.3, pp. 552–568. DOI: 10.1109/TSMCA.2010.2084081 (page 76).

Kim, Ji-Hyun (2009). "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap". *Computational statistics & data analysis* 53.11, pp. 3735–3745 (pages 93, 94).

Kim, Joong Hee et al. (2007). "Noninvasive diffusion tensor imaging of evolving white matter pathology in a mouse model of acute spinal cord injury". en. *Magnetic Resonance in Medicine* 58.2, pp. 253–260 (page 170).

Kimelberg, Harold K (2010). "Functions of mature mammalian astrocytes: a current view". *The Neuroscientist* 16.1, pp. 79–106 (page 19).

Kindermans, P.-J. et al. (2017). "The (Un)reliability of saliency methods". *ArXiv e-prints*. arXiv: 1711.00867 [stat.ML] (page 157).

King, Gary et al. (2001). "Logistic Regression in Rare Events Data". *Polit. Anal.* 9.2, pp. 137–163 (pages 76, 110, 140).

Kingma, Diederik, Ba, Jimmy (2014). "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980* (pages 67, 108, 119, 120).

Kinney, H C et al. (1994). "Myelination in the developing human brain: biochemical correlates." *Neurochemical Research* 19.8, pp. 983–96 (pages 28, 34, 56).

Kinney, H C et al. (1988). "Sequence of central nervous system myelination in human infancy. II. Patterns of myelination in autopsied infants." *Journal of Neuropathology and Experimental Neurology* 47.3, pp. 217–234. DOI: 10.1097/00005072-198805000-00003 (pages 35, 230).

Kjær, Majken et al. (2016). "Neocortical Development in Brain of Young Children—A Stereological Study". *Cerebral Cortex* 27.12, pp. 5477–5484 (page 14).

Klawiter, Eric C et al. (2011). "Radial diffusivity predicts demyelination in ex vivo multiple sclerosis spinal cords". *Neuroimage* 55.4, pp. 1454–1460 (page 169).

Klein, Arno et al. (2009). "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration". *Neuroimage* 46.3, pp. 786–802 (page 190).

Kleinnijenhuis, Michiel et al. (2016). "The effect of axon shape and myelination on diffusion signals in a realistic Monte Carlo simulation environment". *Proc. Int. Soc. Magn. Reson. Med. Sci. Meet. Exhib. Int. Soc. Magn. Reson. Med. Sci. Meet. Exhib.* 23 (page 183).

Klistorner, Alexander et al. (2015). "Decoding diffusivity in multiple sclerosis: analysis of optic radiation lesional and non-lesional white matter". en. *PLoS One* 10.3, e0122114 (page 182).

Knuesel, Irene et al. (2014). "Maternal immune activation and abnormal brain development across CNS disorders". *Nature Reviews Neurology* 10.11, pp. 643–660 (page 17).

Kobayashi, Y, Lavenex, P (2014). "Neuroanatomy Methods in Humans and Animals" (page 27).

Kohavi, Ron et al. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". *Ijcai.* Vol. 14. 2. Montreal, Canada, pp. 1137–1145 (pages 93–95).

Kolind, Shannon et al. (2013). "Myelin imaging in amyotrophic and primary lateral sclerosis". *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 14.7-8, pp. 562–573 (page 169).

Kostović, Ivica, Jovanov-Milošević, Nataša (2006). "The development of cerebral connections during the first 20–45 weeks' gestation". *Seminars in Fetal and Neonatal Medicine*. Vol. 11. 6. Elsevier, pp. 415–422 (pages 17, 24, 28–30).

Kostović, Ivica, Judaš, Miloš (2010). "The development of the subplate and thalamocortical connections in the human foetal brain". *Acta paediatrica* 99.8, pp. 1119–1127 (page 16).

Kostović, Ivica, Judaš, Miloš, Sedmak, Goran (2011). "Developmental history of the subplate zone, subplate neurons and interstitial white matter neurons: relevance for schizophrenia". *International Journal of Developmental Neuroscience* 29.3, pp. 193–205 (page 28).

Kostović, Ivica et al. (2002). "Laminar organization of the human fetal cerebrum revealed by histochemical markers and magnetic resonance imaging". *Cerebral Cortex* 12.5, pp. 536–544 (page 30).

Kotikalapudi, Raghavendra, contributors (2017). *keras-vis*. https://github.com/raghakot/keras-vis (page 163).

Kowalczyk, Tom et al. (2009). "Intermediate neuronal progenitors (basal progenitors) produce pyramidal–projection Neurons for all layers of cerebral cortex". *Cerebral cortex* 19.10, pp. 2439–2450 (page 21).

Kriegstein, Arnold R, Noctor, Stephen C (2004). "Patterns of neuronal migration in the embryonic cortex". *Trends in neurosciences* 27.7, pp. 392–399 (pages 16, 21, 22, 24).

Krizhevsky, Alex, Sutskever, Ilya, Hinton, Geoffrey E (2012). "Imagenet classification with deep convolutional neural networks". *Advances in neural information processing systems*, pp. 1097–1105 (pages 64, 74, 100).

Kucharczyk, Walter et al. (1994). "Relaxivity and magnetization transfer of white matter lipids at MR imaging: importance of cerebrosides and pH." *Radiology* 192.2, pp. 521–529 (page 34).

Kukar, Matjaz, Kononenko, Igor, et al. (1998). "Cost-Sensitive Learning with Neural Networks." *ECAI*, pp. 445–449 (page 76).

Kukačka, J., Golkov, V., Cremers, D. (2017). "Regularization for Deep Learning: A Taxonomy". *ArXiv e-prints*. arXiv: 1710.10686 [cs.LG] (pages 72, 74).

Kuklisova-Murgasova, Maria et al. (2011). "A dynamic 4D probabilistic atlas of the developing brain." *Neuroimage* 54.4, pp. 2750–63. DOI: 10.1016/j.neuroimage.2010.10.019 (page 187).

Kullback, Solomon, Leibler, Richard A (1951). "On information and sufficiency". *The annals of mathematical statistics* 22.1, pp. 79–86 (page 65).

Kumar, Rajesh et al. (2013). "Brain axial and radial diffusivity changes with age and gender in healthy adults". *Brain research* 1512, pp. 22–36 (page 182).

Kunz, Nicolas et al. (2014). "Assessing white matter microstructure of the newborn with multi-shell diffusion MRI and biophysical compartment models". *Neuroimage* 96, pp. 288–299. DOI: 10.1016/j.neuroimage.2014.03.057 (page 230).

Kwok, Jessica CF et al. (2011). "Extracellular matrix and perineuronal nets in CNS repair". *Developmental neurobiology* 71.11, pp. 1073–1089 (page 30).

Kwon, Junhwan et al. (2017). "Label-free nanoscale optical metrology on myelinated axons in vivo". *Nature communications* 8.1, p. 1832 (pages 32, 33).

Lake, Brenden M, Salakhutdinov, Ruslan, Tenenbaum, Joshua B (2015). "Human-level concept learning through probabilistic program induction". *Science* 350.6266, pp. 1332–1338 (page 76).

LaMantia, AS, Rakic, P (1990). "Axon overproduction and elimination in the corpus callosum of the developing rhesus monkey". *Journal of Neuroscience* 10.7, pp. 2156–2175 (pages 30, 31).

Lamblin, Pascal, Bengio, Yoshua (2010). "Important gains from supervised fine-tuning of deep architectures on large labeled sets". *NIPS* 2010 Deep Learning and Unsupervised Feature Learning Workshop* (page 77).

Lampinen, Björn et al. (2017). "Optimal experimental design for filter exchange imaging: Apparent exchange rate measurements in the healthy brain and in intracranial tumors". *Magnetic resonance in medicine* 77.3, pp. 1104–1114 (page 52).

Landman, Bennett A et al. (2010). "Complex geometric models of diffusion and relaxation in healthy and damaged white matter". en. *NMR in Biomedicine* 23.2, pp. 152–162 (page 173).

Langer-Gould, A et al. (2011). "Incidence of acquired CNS demyelinating syndromes in a multiethnic cohort of children." *Neurology* 77.12, pp. 1143–8. DOI: 10.1212/WNL.0b013e31822facdd (page 169).

Larkman, D J et al. (2001). "Use of multicoil arrays for separation of signal from multiple slices simultaneously excited." *Journal of Magnetic Resonance Imaging* 13.2, pp. 313–7 (pages 46, 187).

Lätt, Jimmy et al. (2007a). "Accuracy of q-Space Related Parameters in MRI: Simulations and Phantom Measurements". *IEEE transactions on medical imaging* 26.11, pp. 1437–1447 (page 53).

Lätt, Jimmy et al. (2007b). "Effects of restricted diffusion in a biological phantom: a q-space diffusion MRI study of asparagus stems at a 3T clinical scanner". *Magnetic Resonance Materials in Physics, Biology and Medicine* 20.4, p. 213 (page 53).

Lau, Lorraine W et al. (2013). "Pathophysiology of the brain extracellular matrix: a new target for remyelination". *Nature Reviews Neuroscience* 14.10, nrn3550 (page 30).

Laule, C et al. (2004). "Water content and myelin water fraction in multiple sclerosis: A T 2 relaxation study". *Journal of Neurology* 251.3, pp. 284–293 (page 175).

Lauterbur, P C (1973). "Image Formation by Induced Local Interactions: Examples Employing Nuclear Magnetic Resonance". *Nature* 242.5394, pp. 190–191. DOI: 10.1038/242190a0 (page 40).

Lavesson, Niklas, Davidsson, Paul (2007). "Evaluating learning algorithms and classifiers". *International Journal of Intelligent Information and Database Systems* 1.1, pp. 37–52 (page 79).

Le Bihan, D et al. (2001). "Diffusion tensor imaging: concepts and applications". en. *Journal of Magnetic Resonance Imaging* 13.4, pp. 534–546 (page 54).

Le Bihan, D et al. (1986). "MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders." *Radiology* 161.2, pp. 401–7. DOI: 10.1148/radiology.161.2.3763909 (page 53).

Le Bihan, Denis (1995). "Molecular diffusion, tissue microdynamics and microstructure." *NMR in Biomedicine* 8.7-8, pp. 375–386. DOI: 10.1002/nbm.1940080711 (page 43).

Le Bihan, Denis, Johansen-Berg, Heidi (2012). "Diffusion MRI at 25: exploring brain tissue structure and function". *Neuroimage* 61.2, pp. 324–341 (page 41).

Le Bihan, Denis et al. (2006). "Artifacts and pitfalls in diffusion MRI". *Journal of magnetic resonance imaging* 24.3, pp. 478–488 (pages 46, 98, 169).

Le Bihan, Denis et al. (1986). "MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders." *Radiology* 161.2, pp. 401–407 (page 43).

LeCun, Yann, Bengio, Yoshua, Hinton, Geoffrey (2015). "Deep learning". *Nature* 521.7553, pp. 436–444 (pages 64, 67).

LeCun, Yann, Bengio, Yoshua, et al. (1995). "Convolutional networks for images, speech, and time series". *The handbook of brain theory and neural networks* 3361.10, p. 1995 (page 67).

Lee, Hong-Hsi, Fieremans, Els, Novikov, Dmitry S (2016). "What dominates the time dependence of diffusion transverse to axons: Intra- or extra-axonal water?" arXiv: 1707.09426v1 (page 184).

Lee, June-Goo et al. (2017). "Deep Learning in Medical Imaging: General Overview". *Korean Journal of Radiology* 18.4, p. 570. DOI: 10.3348/kjr.2017.18.4.570 (page 64).

Leemans, Alexander, Jones, Derek K (2009). "The B-matrix must be rotated when correcting for subject motion in DTI data". *Magnetic resonance in medicine* 61.6, pp. 1336–1349 (pages 47, 98).

Leergaard, Trygve B et al. (2010). "Quantitative histological validation of diffusion MRI fiber orientation distributions in the rat brain". en. *PLoS One* 5.1, e8595 (page 176).

Lehtinen, Maria K, Walsh, Christopher A (2011). "Neurogenesis at the brain–cerebrospinal fluid interface". *Annual review of cell and developmental biology* 27, pp. 653–679 (page 21).

Leipsic, Paul Flechsig Of (1901). "Developmental (myelogenetic) localisation of the cerebral cortex in the human subject." *The Lancet* 158.4077, pp. 1027–1030 (page 35).

Leonard, Christiana M et al. (1998). "Normal variation in the frequency and location of human auditory cortex landmarks. Heschl's gyrus: where is it?" *Cerebral cortex (New York, NY: 1991)* 8.5, pp. 397–406 (page 199).

Letourneau, Paul C, Condic, Maureen L, Snow, Diane M (1994). "Interactions of Developing Neurons with the Extracellular Matrix". *The Journal of Neuroscience* 14.3, pp. 915–928 (page 30).

Li, X et al. (2018). "Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift". *ArXiv e-prints*. arXiv: 1801.05134 [cs.LG] (page 126).

Li, Yue et al. (2013). "Image corruption detection in diffusion tensor imaging for post-processing and real-time monitoring". *PloS one* 8.10, e49764 (page 99).

Lian, J, Williams, D S, Lowe, I J (1994). "Magnetic Resonance Imaging of Diffusion in the Presence of Background Gradients and Imaging of Background Gradients". *Journal of Magnetic Resonance. Series A* 106.1, pp. 65–74 (page 183).

Liewald, Daniel et al. (2014). "Distribution of axon diameters in cortical white matter: an electron-microscopic study on three human brains and a macaque". *Biological Cybernetics* (pages 174, 175, 183).

Lin, Shih-chun, Bergles, Dwight E (2004). "Synaptic signaling between GABAergic interneurons and oligodendrocyte precursor cells in the hippocampus". *Nature neuroscience* 7.1, p. 24 (page 33).

Linderkamp, Otwin et al. (2009). "Time table of normal foetal brain development". *International Journal of Prenatal and Perinatal Psychology and Medicine* 21.1/2, pp. 4–16 (page 17).

Ling, Charles X, Huang, Jin, Zhang, Harry, et al. (2003). "AUC: a statistically consistent and more discriminating measure than accuracy". *IJCAI*. Vol. 3, pp. 519–524 (pages 80, 82).

Ling, Charles X, Zhang, Huajie (2002). "Toward Bayesian classifiers with accurate probabilities". *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 123–134 (page 82).

Ling, Josef et al. (2012). "Head injury or head motion? Assessment and quantification of motion artifacts in diffusion tensor imaging studies". *Human brain mapping* 33.1, pp. 50–62 (page 99).

Litjens, Geert et al. (2017). "A survey on deep learning in medical image analysis". *arXiv preprint arXiv:1702.05747* (page 64).

Liu, Bilan, Zhu, Tong, Zhong, Jianhui (2015). "Comparison of quality control software tools for diffusion tensor imaging". *Magnetic resonance imaging* 33.3, pp. 276–285 (page 99).

Liu, H. et al. (2017). "Hierarchical Representations for Efficient Architecture Search". *ArXiv e-prints*. arXiv: 1711.00436 [cs.LG] (page 152).

Liu, Wei et al. (2013). "Slice-wise optimization algorithm for diffusion tensor estimation". *Dongnan Daxue Xuebao (Ziran Kexue Ban)/Journal of Southeast University (Natural Science Edition)* 43.1, pp. 30–34 (page 100).

Llinás, Rodolfo R (2003). "The contribution of Santiago Ramon y Cajal to functional neuroscience". *Nature Reviews Neuroscience* 4.1, pp. 77–80 (page 14).

Lodygensky, G A et al. (2010). "In vivo MRI analysis of an inflammatory injury in the developing brain". en. *Brain Behav. Immun.* 24.5, pp. 759–767 (page 170).

Lohmann, Gabriele, Cramon, D Yves von, Steinmetz, Helmuth (1999). "Sulcal variability of twins". *Cerebral Cortex* 9.7, pp. 754–763 (page 199).

López, Victoria, Fernández, Alberto, Herrera, Francisco (2014). "On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed". *Information Sciences* 257, pp. 1–13 (page 79).

López, Victoria et al. (2013). "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics". *Inf. Sci. (Ny).* 250, pp. 113–141. DOI: 10.1016/j.ins.2013.07.007 (pages 74, 76).

López, Victoria et al. (2012). "Cost Sensitive and Preprocessing for Classification with Imbalanced Data-sets: Similar Behaviour and Potential Hybridizations." *ICPRAM (2)*, pp. 98–107 (page 76).

Lorch, Benedikt et al. (2017). "Automated Detection of Motion Artefacts in MR Imaging Using Decision Forests". *Journal of medical engineering* 2017 (page 100).

Lorenzen, Peter, Davis, Brad C, Joshi, Sarang (2005). "Unbiased atlas formation via large deformations metric mapping". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 411–418 (page 192).

Lorenzen, Peter et al. (2006). "Multi-modal image set registration and atlas formation". *Medical image analysis* 10.3, pp. 440–451 (page 193).

Lu, Le et al. (2016). "Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer Learning Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer". *IEEE Transactions on Medical Imaging* 35.5, pp. 1285–1298. DOI: 10.1109/TMI.2016.2528162. arXiv: 1602.03409 (pages 94, 101).

Lucic, M. et al. (2017). "Are GANs Created Equal? A Large-Scale Study". *ArXiv e-prints*. arXiv: 1711.10337 [stat.ML] (page 94).

Mackay, Alex et al. (1994). "In vivo visualization of myelin water in brain by magnetic resonance". *Magnetic resonance in medicine* 31.6, pp. 673–677 (page 49).

Mackay, Alex et al. (2006). "Insights into brain microstructure from the T2 distribution". *Magnetic resonance imaging* 24.4, pp. 515–525 (page 49).

Mackenzie, I S et al. (2014). "Incidence and prevalence of multiple sclerosis in the UK 1990–2010: a descriptive study in the General Practice Research Database". *Journal of Neurology, Neurosurgery and Psychiatry* 85, pp. 76–84. DOI: 10.1136/jnnp-2013-305450 (page 169).

Mahad, Don H, Trapp, Bruce D, Lassmann, Hans (2015). "Pathological mechanisms in progressive multiple sclerosis". *Lancet Neurology* 14.2, pp. 183–193 (page 170).

Makropoulos, Antonios et al. (2014). "Automatic whole brain MRI segmentation of the developing neonatal brain". *IEEE transactions on medical imaging* 33.9, pp. 1818–1831 (page 204).

Mangin, J-F et al. (2002). "Distortion correction and robust tensor estimation for MR diffusion imaging". *Medical Image Analysis* 6.3, pp. 191–198 (page 99).

Mansfield, Peter (1977). "Multi-planar image formation using NMR spin echoes". *Journal of Physics C: Solid State Physics* 10.3, p. L55 (page 44).

Marín, Oscar, Rubenstein, John LR (2001). "A long, remarkable journey: tangential migration in the telencephalon". *Nature Reviews Neuroscience* 2.11, pp. 780–790 (page 24).

Martin, C. H., Mahoney, M. W. (2017). "Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior". *ArXiv e-prints*. arXiv: 1710.09553 [cs.LG] (page 73).

Masland, Richard H (2004). "Neuronal cell types". *Current Biology* 14.13, R497–R500 (page 18).

Mason, Simon J, Graham, Nicholas E (2002). "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation". *Quarterly Journal of the Royal Meteorological Society* 128.584, pp. 2145–2166 (page 81).

Matsumae, Mitsunori et al. (2001). "Sequential changes in MR water proton relaxation time detect the process of rat brain myelination during maturation". *Mechanisms of Ageing and Development* 122.12, pp. 1281–1291. DOI: 10.1016/S0047-6374(01)00265-2 (page 34).

Matthews, Brian W (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2, pp. 442–451 (page 80).

McArdle, CB et al. (1987). "Developmental features of the neonatal brain: MR imaging. Part I. Gray-white matter differentiation and myelination." *Radiology* 162.1, pp. 223–229 (page 35).

McConnell, Susan K, Ghosh, Anirvan, Shatz, Carla J (1989). "Subplate neurons pioneer the first axon pathway from the cerebral cortex". *Science* 245.4921, pp. 978–982 (page 28).

McGibney, G et al. (1993). "Quantitative evaluation of several partial Fourier reconstruction algorithms used in MRI". *Magnetic resonance in medicine* 30.1, pp. 51–59 (page 45).

McKinstry, Robert C et al. (2002). "Radial organization of developing preterm human cerebral cortex revealed by non-invasive water diffusion anisotropy MRI". *Cerebral Cortex* 12.12, pp. 1237–1243 (pages 206, 212, 225).

McTigue, Dana M, Tripathi, Richa B (2008). "The life, death, and replacement of oligodendrocytes in the adult CNS". *Journal of neurochemistry* 107.1, pp. 1–19 (pages 31, 33).

Mehta, P., Schwab, D. J. (2014). "An exact mapping between the Variational Renormalization Group and Deep Learning". *ArXiv e-prints*. arXiv: 1410.3831 [stat.ML] (page 120).

Melbourne, Andrew et al. (2014). "Multi-modal measurement of the myelin-to-axon diameter g-ratio in preterm-born neonates and adult controls". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 268–275 (page 184).

Menegola, A. et al. (2016). "Towards Automated Melanoma Screening: Exploring Transfer Learning Schemes". *ArXiv e-prints*. arXiv: 1609.01228 [cs.CV] (page 77).

Michalski, John-Paul, Kothary, Rashmi (2015). "Oligodendrocytes in a Nutshell". *Frontiers in cellular neuroscience* 9, p. 340 (pages 33, 34).

Mikula, Shawn, Binding, Jonas, Denk, Winfried (2012). "Staining and embedding the whole mouse brain for electron microscopy". *Nature methods* 9.12, p. 1198 (pages 50, 51).

Mikula, Shawn, Denk, Winfried (2015). "High-resolution whole-brain staining for electron microscopic circuit reconstruction". *nAture methods* 12.6, p. 541 (page 50).

Miller, Karla L et al. (2011). "Diffusion imaging of whole, post-mortem human brains on a clinical MRI scanner". *Neuroimage* 57.1, pp. 167–181 (pages 199, 207, 225).

Miller, Robert H (2002). "Regulation of oligodendrocyte development in the vertebrate CNS". *Progress in neurobiology* 67.6, pp. 451–467 (page 33).

Minati, Ludovico, Węglarz, Władysław P (2007). "Physical foundations, models, and methods of diffusion magnetic resonance imaging of the brain: A review". *Concepts in Magnetic Resonance Part A* 30.5, pp. 278–307 (page 41).

Molinaro, Annette M, Simon, Richard, Pfeiffer, Ruth M (2005). "Prediction error estimation: a comparison of resampling methods". *Bioinformatics* 21.15, pp. 3301–3307 (page 94).

Mollink, Jeroen et al. (2017). "Evaluating fibre orientation dispersion in white matter: Comparison of diffusion MRI, histology and polarized light imaging". *Neuroimage* 157, pp. 561–574 (pages 50, 52).

Molliver, Mark E, Kostovic, Ivica, Loos, Hendrik Van der (1973). "The development of synapses in cerebral cortex of the human fetus". *Brain research* 50.2, pp. 403–407 (page 16).

Montúfar, G. et al. (2014). "On the Number of Linear Regions of Deep Neural Networks". *ArXiv e-prints*. arXiv: 1402.1869 [stat.ML] (page 71).

Mor-Yosef, S et al. (1990). "Ranking the risk factors for cesarean: logistic regression analysis of a nationwide study." *Obstetrics and Gynecology* 75.6, pp. 944–7 (page 66).

Mori, Susumu et al. (2008). "Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template". *Neuroimage* 40.2, pp. 570–582 (page 187).

Mori, Susumu et al. (1999). "Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging". *Annals of neurology* 45.2, pp. 265–269 (page 56).

Moseley, Michael E et al. (1990). "Diffusion-weighted MR imaging of anisotropic water diffusion in cat central nervous system." *Radiology* 176.2, pp. 439–445 (page 48).

Mottershead, J P et al. (2003). "High field MRI correlates of myelin content and axonal density in multiple sclerosis: A post-mortem study of the spinal cord". *Journal of Neurology* 250.11, pp. 1293–1301 (pages 170, 182).

Mrzljak, Ladislav et al. (1988). "Prenatal development of neurons in the human prefrontal cortex: I. A qualitative Golgi study". *Journal of comparative neurology* 271.3, pp. 355–386 (pages 17, 22, 27).

Myelinated neuron (2018). *File:Myelinated_ neuron.jpg — Wikipedia, The Free Encyclopedia*. [Online; accessed 23-February-2018] (page 32).

Nave, Klaus-Armin (2010). "Myelination and support of axonal integrity by glia". *Nature* 468.7321, p. 244 (page 19).

Nedjati-Gilani, Gemma L et al. (2017). "Machine learning based compartment models with permeability for white matter microstructure imaging". *Neuroimage* 150, pp. 119–135 (page 57).

Neil, J. et al. (2002). *Diffusion tensor imaging of normal and injured developing human brain - A technical review*. DOI: 10.1002/nbm.784 (page 169).

Ng, Andrew (2012). "CS229 Lecture notes - Supervised learning" (pages 65, 66).

Nguyen, A, Yosinski, J, Clune, J (2014). "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images". *ArXiv e-prints*. arXiv: 1412.1897 [cs.CV] (page 156).

Nguyen, Q, Hein, M (2017). "The loss surface of deep and wide neural networks". *ArXiv e-prints*. arXiv: 1704.08045 [cs.LG] (page 72).

Nicholson, Charles, Syková, Eva (1998). "Extracellular space structure revealed by diffusion analysis". *Trends in neurosciences* 21.5, pp. 207–215 (page 30).

Nilsson, Markus et al. (2013a). "Noninvasive mapping of water diffusional exchange in the human brain using filter-exchange imaging". *Magnetic resonance in medicine* 69.6, pp. 1572–1580 (page 52).

Nilsson, Markus et al. (2009). "On the effects of a varied diffusion time in vivo: is the diffusion in white matter restricted?" *Magnetic resonance imaging* 27.2, pp. 176–187 (page 52).

Nilsson, Markus et al. (2017). "Resolution limit of cylinder diameter estimation by diffusion MRI: The impact of gradient waveform and orientation dispersion". *NMR in Biomedicine* 30.7 (page 52).

Nilsson, Markus et al. (2013b). "The role of tissue microstructure and water exchange in biophysical modelling of diffusion in white matter". *Magnetic Resonance Materials in Physics, Biology and Medicine* 26.4, pp. 345–370 (pages 52, 53).

Norris, David G (2001). "Implications of bulk motion for diffusion-weighted imaging experiments: Effects, mechanisms, and solutions". *Journal of magnetic resonance imaging* 13.4, pp. 486–495 (page 47).

Novak, Ulrike, Kaye, Andrew H (2000). "Extracellular matrix and the brain: components and function". *Journal of clinical neuroscience* 7.4, pp. 280–290 (page 30).

Novikov, Dmitry S, Fieremans, Els (2012). "Relating extracellular diffusivity to cell size distribution and packing density as applied to white matter". *Proceedings of the 20th Annual Meeting of ISMRM, Melbourne, Victoria, Australia*, p. 1829 (page 58).

Novikov, Dmitry S., Kiselev, Valerij G., Jespersen, Sune N. (2018). "On modeling". *Magnetic Resonance in Medicine* 79.6, pp. 3172–3193. DOI: 10.1002/mrm.27101 (pages 11, 55, 236).

Novikov, Dmitry S et al. (2016). "Quantifying brain microstructure with diffusion MRI: Theory and parameter estimation". *arXiv preprint arXiv:1612.02059* (pages 52, 55, 56, 58).

O'Connor, K M et al. (2013). ""Dazed and diffused": making sense of diffusion abnormalities in neurologic pathologies." *British Journal of Radiology* 86.1032, p. 20130599. DOI: 10.1259/bjr.20130599 (page 170).

Oishi, Kenichi et al. (2011). "Multi-contrast human neonatal brain atlas: Application to normal neonate development analysis". *Neuroimage* 56.1, pp. 8–20. DOI: 10.1016/j.neuroimage.2011.01.051 (page 187).

Olah, Chris, Mordvintsev, Alexander, Schubert, Ludwig (2017). "Feature Visualization". *Distill* 2.11, e7 (pages 70, 157).

Oliveira, Francisco PM, Tavares, Joao Manuel RS (2014). "Medical image registration: a review". *Computer methods in biomechanics and biomedical engineering* 17.2, pp. 73–93 (page 187).

O'Rahilly, Ronan, Müller, Fabiola (2010). "Developmental stages in human embryos: revised and new measurements". *Cells, Tissues, Organs* 192.2, pp. 73–84 (page 14).

– (2000). "Prenatal ages and stages—measures and errors". *Teratology* 61.5, pp. 382–384 (page 15).

– (2006). *The embryonic human brain: an atlas of developmental stages*. John Wiley & Sons (pages 14, 15, 17, 28).

Ourselin, Sébastien et al. (2001). "Reconstructing a 3D structure from serial histological sections". *Image and vision computing* 19.1-2, pp. 25–31 (page 193).

Paass, Gerhard (1993). "Assessing and improving neural network predictions by the bootstrap algorithm". *Advances in Neural Information Processing Systems*, pp. 196–203 (page 95).

Pan, Sinno Jialin, Yang, Qiang (2010). "A survey on transfer learning". *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359 (page 77).

Panagiotaki, Eleftheria, Hall, M G, Zhang, Hui (2010). "High-fidelity meshes from tissue samples for diffusion MRI simulations". *Image Computing and*, pp. 404–411 (page 183).

Panagiotaki, Eleftheria et al. (2012). "Compartment models of the diffusion MR signal in brain white matter: a taxonomy and comparison". *Neuroimage* 59.3, pp. 2241–2254 (pages 57, 172).

Pannek, Kerstin et al. (2017). "Automatic detection of volumes affected by subvolume movement". *ISMRM 25th Annu. Meet. Exhib.* (Page 99).

Pannek, Kerstin et al. (2012a). "Diffusion MRI of the neonate brain: acquisition, processing and analysis techniques". *Pediatric radiology* 42.10, pp. 1169–1182 (page 99).

Pannek, Kerstin et al. (2018). "Fixel-based analysis reveals alterations is brain microstructure and macrostructure of preterm-born infants at term equivalent age". *NeuroImage: Clinical* (page 237).

Pannek, Kerstin et al. (2012b). "HOMOR: higher order model outlier rejection for high b-value MR diffusion data". *Neuroimage* 63.2, pp. 835–842 (page 99).

Paola, JD, Schowengerdt, RA (1995). "A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery". *International Journal of remote sensing* 16.16, pp. 3033–3058 (page 72).

Papadopoulos, Marios C, Verkman, Alan S (2013). "Aquaporin water channels in the nervous system". *Nature Reviews Neuroscience* 14.4, p. 265 (page 52).

Parekh, Ruchi, Ascoli, Giorgio A (2013). "Neuronal morphology goes digital: a research hub for cellular and system neuroscience". *Neuron* 77.6, pp. 1017–1038 (page 20).

Park, Hyunjin et al. (2005). "Least biased target selection in probabilistic atlas construction". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 419–426 (page 192).

Park, Laurence A (2011). "Bootstrap confidence intervals for mean average precision". *Proceedings of the 4th Applied Statistics Education and Research Collaboration (ASEARC) Conference, Paramatta, February 17-18, 2011*, pp. 51–54 (page 86).

Parker, Charles (2013). "On measuring the performance of binary classifiers". *Knowledge and information systems* 35.1, pp. 131–152 (pages 79, 80, 82, 83, 92).

Parker, Sara S et al. (2013). "Competing molecular interactions of aPKC isoforms regulate neuronal polarity". *Proceedings of the National Academy of Sciences* 110.35, pp. 14450–14455 (page 18).

Parra, Paula, Gulyas, Attila I, Miles, Richard (1998). "How many subtypes of inhibitory cells in the hippocampus?" *Neuron* 20.5, pp. 983–993 (page 18).

Pascanu, Razvan, Montufar, Guido, Bengio, Yoshua (2013). "On the number of response regions of deep feed forward networks with piece-wise linear activations". *arXiv preprint arXiv:1312.6098* (page 71).

Patel, Vishal et al. (2010). "LONI MiND: metadata in NIfTI for DWI". *Neuroimage* 51.2, pp. 665–676 (page 187).

Pavaine, Julia et al. (2016). "Diffusion tensor imaging-based assessment of white matter tracts and visual-motor outcomes in very preterm neonates". *Neuroradiology* 58.3, pp. 301–310 (page 99).

Pecheva, Diliana et al. (2017). "White matter diffusion properties at term equivalent age are associated with subsequent motor performance in infants born preterm". *ISMRM 25th Annu. Meet. Exhib.* (Page 237).

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine learning in Python". *Journal of machine learning research* 12.Oct, pp. 2825–2830 (page 80).

Peng, Huiling et al. (2009). "Development of a human brain diffusion tensor template". *Neuroimage* 46.4, pp. 967–980 (page 187).

Pereyra, G. et al. (2017). "Regularizing Neural Networks by Penalizing Confident Output Distributions". *ArXiv e-prints*. arXiv: 1701.06548 (page 74).

Perez, L., Wang, J. (2017). "The Effectiveness of Data Augmentation in Image Classification using Deep Learning". *ArXiv e-prints*. arXiv: 1712.04621 [cs.CV] (pages 104, 138, 152).

Perge, János A et al. (2009). "How the optic nerve allocates space, energy capacity, and information". *Journal of Neuroscience* 29.24, pp. 7917–7928 (page 52).

Perrin, J. S. et al. (2009). "Sex differences in the growth of white matter during adolescence". *Neuroimage* 45.4, pp. 1055–1066. DOI: 10.1016/j.neuroimage.2009.01.023 (page 169).

Pfeuffer, Josef, Provencher, Stephen W, Gruetter, Rolf (1999). "Water diffusion in rat brain in vivo as detected at very large b values is multicompartmental". *Magnetic Resonance Materials in Physics, Biology and Medicine* 8.2, pp. 98–108 (page 52).

Pfeuffer, Josef et al. (1998). "Expression of aquaporins in Xenopus laevis oocytes and glial cells as detected by diffusion-weighted 1H NMR spectroscopy and photometric swelling assay". *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1448.1, pp. 27–36 (page 52).

Philibert, Jean (2005). "One and a half century of diffusion: Fick, Einstein, before and beyond". *Diffusion Fundamentals* 2.1, pp. 1–10 (page 37).

Picard, Richard R, Berk, Kenneth N (1990). "Data splitting". *The American Statistician* 44.2, pp. 140–147 (page 92).

Pietsch, Maximilian, Tournier, J-Donald (2015). "Effect of demyelination on diffusion tensor indices: a Monte Carlo simulation study". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 23, p. 3039 (page 11).

Pietsch, Maximilian et al. (2018). "Longitudinal multi-component HARDI atlas of neonatal white matter". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 26, p. 0469 (page 11).

Pietsch, Maximilian et al. (2017a). "Multi-contrast diffeomorphic non-linear registration of orientation density functions". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 25 (pages 11, 187, 194, 221).

Pietsch, Maximilian et al. (2017b). "Multi-shell neonatal brain HARDI template". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 25 (pages 11, 187, 216).

Poggio, T. et al. (2018). "Theory of Deep Learning III: explaining the non-overfitting puzzle". *ArXiv e-prints*. arXiv: 1801.00173 [cs.LG] (page 72).

Poggio, Tomaso (2011). "The computational magic of the ventral stream" (page 70).

Pollock, Jonathan D, Wu, Da-Yu, Satterlee, John S (2014). "Molecular neuroanatomy: a generation of progress." *Trends in Neurosciences* 37.2, pp. 106–23. DOI: 10.1016/j.tins.2013.11.001 (page 56).

Poncelet, BP et al. (1992). "Brain parenchyma motion: measurement with cine echo-planar MR imaging." *Radiology* 185.3, pp. 645–651 (page 48).

Pontius Jr, Robert Gilmore, Millones, Marco (2011). "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment". *International Journal of Remote Sensing* 32.15, pp. 4407–4429 (page 80).

Powers, David Martin (2011). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation" (page 80).

– (2012). "The problem with kappa". *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 345–355 (pages 80, 81).

Price, William S (1997). "Pulsed-field gradient nuclear magnetic resonance as a tool for studying translational diffusion: Part 1. Basic theory". *Concepts in Magnetic Resonance Part A* 9.5, pp. 299–336 (page 41).

Pruessmann, Klaas P et al. (1999). "SENSE: sensitivity encoding for fast MRI". *Magnetic resonance in medicine* 42.5, pp. 952–962 (page 45).

Pul, Carola van et al. (2012). "Quantitative fiber tracking in the corpus callosum and internal capsule reveals microstructural abnormalities in preterm infants at term-equivalent age". *American Journal of Neuroradiology* 33.4, pp. 678–684 (page 99).

Purcell, Edward M, Torrey, H Co, Pound, Robert V (1946). "Resonance absorption by nuclear magnetic moments in a solid". *Physical review* 69.1-2, p. 37 (page 39).

Quarles, Richard H, Macklin, Wendy B, Morell, Pierre (2006). "Myelin formation, structure and biochemistry". *Basic neurochemistry: molecular, cellular and medical aspects* 7, pp. 51–71 (pages 31, 34).

Quirk, James D et al. (2003). "Equilibrium water exchange between the intra-and extracellular spaces of mammalian brain". *Magnetic resonance in medicine* 50.3, pp. 493–499 (pages 52, 53, 184).

Rabinowicz, Theodore et al. (1996). "Human cortex development: estimates of neuronal numbers indicate major loss late during gestation". *Journal of neuropathology and experimental neurology* 55.3, pp. 320–328 (page 30).

Raff, Martin C et al. (1993). "Programmed cell death and the control of cell survival: lessons from the nervous system". *Science* 262.5134, pp. 695–700 (page 33).

Raffelt, David et al. (2012). "Apparent Fibre Density: a novel measure for the analysis of diffusion-weighted magnetic resonance images." *Neuroimage* 59.4, pp. 3976–3994 (pages 59, 184, 191).

Raffelt, David et al. (2017). "Bias Field Correction and Intensity Normalisation for Quantitative Analysis of Apparent Fibre Density". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 25, p. 3541 (pages 196, 220).

Raffelt, David et al. (2011). "Symmetric diffeomorphic registration of fibre orientation distributions". *Neuroimage* 56.3, pp. 1171–1180. DOI: 10.1016/j.neuroimage.2011.02.014 (pages 191–194, 199, 220, 221).

Raffelt, David A et al. (2015). "Connectivity-based fixel enhancement: Whole-brain statistical analysis of diffusion MRI measures in the presence of crossing fibres". *Neuroimage* 117, pp. 40–55 (page 56).

Raffelt, David A et al. (2016). "Investigating White Matter Fibre Density and Morphology using Fixel-Based Analysis". *Neuroimage*. DOI: 10.1016/j.neuroimage.2016.09.029 (page 232).

Rakic, Pasko (2003). "Developmental and evolutionary adaptations of cortical radial glia". *Cerebral cortex* 13.6, pp. 541–549 (pages 20, 24).

– (2009). "Evolution of the neocortex: a perspective from developmental biology". *Nature Reviews Neuroscience* 10.10, pp. 724–735 (page 21).

– (1990). "Principles of neural cell migration". *Cellular and Molecular Life Sciences* 46.9, pp. 882–891 (pages 16, 20–22, 24).

– (1988). "Specification of cerebral cortical areas". *Science* 241.4862, p. 170 (page 26).

Rakic, Sonja, Zecevic, Nada (2000). "Programmed cell death in the developing human telencephalon". *European Journal of Neuroscience* 12.8, pp. 2721–2734 (page 30).

Rand, RP, Fuller, NL, Lis, LJ (1979). "Myelin swelling and measurement of forces between myelin membranes". *Nature* 279.5710, pp. 258–260 (page 33).

Rasul, Kashif (2017). *Github gist*. https://gist.github.com/kashif/76792939dd6f473b7404474989cb62a8/dbbfe8444d6712ca988fc12184b61f06fcf65ceb (page 106).

Readhead, C, Hood, L (1990). "The dysmyelinating mouse mutations shiverer (shi) and myelin deficient (shimld)". *Behavior Genetics* 20.2, pp. 213–234 (page 182).

Real, Esteban et al. (2017). "Large-scale evolution of image classifiers". *arXiv preprint arXiv:1703.01041* (page 152).

Recht, B., Re, C. (2012). "Beneath the valley of the noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences". *ArXiv e-prints*. arXiv: 1202.4184 [math.OC] (page 105).

Reisert, Marco et al. (2017). "Disentangling micro from mesostructure by diffusion MRI: A Bayesian approach". *Neuroimage* 147, pp. 964–975 (pages 56, 58).

Reuter, Martin et al. (2012). "Within-subject template estimation for unbiased longitudinal image analysis". *Neuroimage* 61.4, pp. 1402–1418 (page 192).

Reynolds, Richard, Hardy, Rebecca (1997). "Oligodendroglial progenitors labeled with the O4 antibody persist in the adult rat cerebral cortex in vivo". *Journal of neuroscience research* 47.5, pp. 455–470 (page 33).

Riccomagno, Martin M, Kolodkin, Alex L (2015). "Sculpting neural circuits by axon and dendrite pruning". *Annual review of cell and developmental biology* 31, pp. 779–805 (page 30).

Richardson, William D, Kessaris, Nicoletta, Pringle, Nigel (2006). "Oligodendrocyte wars". *Nature Reviews Neuroscience* 7.1, p. 11 (pages 25, 31).

Ridgway, Gerard R et al. (2009). "Issues with threshold masking in voxel-based morphometry of atrophied brains". *Neuroimage* 44.1, pp. 99–111 (page 61).

Riesenhuber, Maximilian, Poggio, Tomaso (1999). "Hierarchical models of object recognition in cortex". *Nature neuroscience* 2.11, p. 1019 (page 67).

Robert, Christian (2014). *Machine learning, a probabilistic perspective* (page 64).

Rohlfing, Torsten (2012). "Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable". *IEEE transactions on medical imaging* 31.2, pp. 153–163 (page 189).

Roine, Timo et al. (2014). "Isotropic non-white matter partial volume effects in constrained spherical deconvolution". *Frontiers in neuroinformatics* 8, p. 28 (page 60).

Rolls, Edmund (2013). "The mechanisms for pattern completion and pattern separation in the hippocampus". *Frontiers in systems neuroscience* 7, p. 74 (page 18).

Ronan, Lisa, Fletcher, Paul C (2015). "From genes to folds: a review of cortical gyrification theory". *Brain Structure and Function* 220.5, pp. 2475–2483 (page 199).

Ruder, Sebastian (2016). "An overview of gradient descent optimization algorithms". *arXiv preprint arXiv:1609.04747* (page 67).

Ruder, Sebastian (2017). "An Overview of Multi-Task Learning in Deep Neural Networks". *ArXiv e-prints*. arXiv: 1706.05098 [cs.LG] (page 115).

Ruder, Sebastian, Plank, Barbara (2017). "Learning to select data for transfer learning with Bayesian Optimization". *arXiv preprint arXiv:1707.05246* (page 77).

Rueckert, Daniel et al. (2006). "Diffeomorphic registration using B-splines". *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 702–709 (page 190).

Rumelhart, David E, Hinton, Geoffrey E, Williams, Ronald J (1985). *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science (page 72).

– (1986). "Learning representations by back-propagating errors". *nature* 323.6088, p. 533 (page 72).

Rushton, W A H (1951). "a Theory of the Effects of Fibre Size in Medullated Nerve". *J. Physiol.* 5.1, pp. 101–122. DOI: 10.1113/jphysiol.1951.sp004655 (page 169).

Russakovsky, Olga et al. (2015). "ImageNet Large Scale Visual Recognition Challenge". *Int J Comput Vis* 115, pp. 211–252. DOI: 10.1007/s11263-015-0816-y (page 64).

Sabour, S., Frosst, N., E Hinton, G. (2017). "Dynamic Routing Between Capsules". *ArXiv e-prints*. arXiv: 1710.09829 [cs.CV] (page 70).

Saerens, Marco, Latinne, Patrice, Decaestecker, Christine (2002). "Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure". *Neural computation* 14.1, pp. 21–41 (page 76).

Saffman, PG, Delbrück, M (1975). "Brownian motion in biological membranes". *Proceedings of the National Academy of Sciences* 72.8, pp. 3111–3113 (page 49).

Samsonov, Alexey et al. (2012). "Quantitative MR imaging of two-pool magnetization transfer model parameters in myelin mutant shaking pup". *Neuroimage* 62.3, pp. 1390–1398. DOI: 10.1016/j.neuroimage.2012.05.077 (page 175).

Sanai, Nader et al. (2011). "Corridors of migrating neurons in the human brain and their decline during infancy". *Nature* 478.7369, p. 382 (pages 16, 24).

Sánchez, Ivelisse et al. (1996). "Oligodendroglia regulate the regional expansion of axon caliber and local accumulation of neurofilaments during development independently of myelin formation". *Journal of Neuroscience* 16.16, pp. 5095–5105 (page 31).

Schain, Aaron J, Hill, Robert A, Grutzendler, Jaime (2014). "Label-free in vivo imaging of myelinated axons in health and disease with spectral confocal reflectance microscopy". *Nature medicine* 20.4, p. 443 (page 33).

Schilling, Kurt et al. (2018). "Confirmation of a gyral bias in diffusion MRI fiber tractography". *Human brain mapping* 39.3, pp. 1449–1466 (page 199).

Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". *Neural networks* 61, pp. 85–117 (pages 64, 67).

Schröder, J. M., Bohl, J., Bardeleben, U. von (1988). "Changes of the ratio between myelin thickness and axon diameter in human developing sural, femoral, ulnar, facial, and trochlear nerves". *Acta Neuropathol.* 76.5, pp. 471–483. DOI: 10.1007/BF00686386 (page 169).

Schuh, Andreas et al. (2014). "Construction of a 4D Brain Atlas and Growth Model Using Diffeomorphic Registration". *STIA 2012 MICCAI 2012 Work.* Lecture Notes in Computer Science 8682, pp. 27–37. DOI: 10.1007/978-3-319-14905-9 (page 187).

Serag, Ahmed et al. (2012a). "A multi-channel 4D probabilistic atlas of the developing brain: application to fetuses and neonates". *Annals of the BMVA* 2012.3, pp. 1–14 (page 187).

Serag, Ahmed et al. (2012b). "Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression". *Neuroimage* 59.3, pp. 2255–2265. DOI: 10.1016/j.neuroimage.2011.09.062 (page 187).

Setsompop, Kawin et al. (2012). "Blipped-controlled aliasing in parallel imaging for simultaneous multislice echo planar imaging with reduced g-factor penalty". *Magnetic resonance in medicine* 67.5, pp. 1210–1224 (page 46).

Sharif Razavian, A. et al. (2014). "CNN Features off-the-shelf: an Astounding Baseline for Recognition". *ArXiv e-prints.* arXiv: 1403.6382 [cs.CV] (page 101).

Sherbondy, Anthony J, Rowe, Matthew C, Alexander, Daniel C (2010). "LNCS 6361 - MicroTrack: An Algorithm for Concurrent Projectome and Microstructure Estimation" (page 232).

Sherman, Diane L, Brophy, Peter J (2005). "Mechanisms of axon ensheathment and myelin growth". *Nature Reviews Neuroscience* 6.9, p. 683 (pages 33, 34, 169).

Sherman, Larry S, Back, Stephen A (2008). "A 'GAG'reflex prevents repair of the damaged CNS". *Trends in neurosciences* 31.1, pp. 44–52 (page 30).

Shi, Feng et al. (2014). "Neonatal atlas construction using sparse representation". *Human Brain Mapping* 35.9, pp. 4663–4677. DOI: 10.1002/hbm.22502 (page 187).

Shi, Leming et al. (2010). "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models". *Nature biotechnology* 28.8, p. 827 (pages 80, 92, 95).

Shwartz-Ziv, R., Tishby, N. (2017). "Opening the Black Box of Deep Neural Networks via Information". *ArXiv e-prints*. arXiv: 1703.00810 [cs.LG] (pages 73, 120).

Sidman, Richard L, Rakic, Pasko (1973). "Neuronal migration, with special reference to developing human brain: a review". *Brain research* 62.1, pp. 1–35 (pages 16, 24).

Simonyan, K., Vedaldi, A., Zisserman, A. (2013). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". *ArXiv e-prints*. arXiv: 1312.6034 [cs.CV] (page 163).

Simonyan, Karen, Zisserman, Andrew (2014). "Very deep convolutional networks for large-scale image recognition". *arXiv preprint arXiv:1409.1556* (pages 74, 102, 120).

Sinnaeve, Davy (2012). "The Stejskal–Tanner equation generalized for any gradient shape—an overview of most pulse sequences measuring free diffusion". *Concepts in Magnetic Resonance Part A* 40.2, pp. 39–65 (page 43).

Sloan, Peter-Pike (2008). "Stupid spherical harmonics (sh) tricks". *Game developers conference*. Vol. 9. Citeseer (page 60).

Smith, Stephen M (2002). "Fast robust automated brain extraction". *Human Brain Mapping* 17.3, pp. 143–155. DOI: 10.1002/hbm.10062 (page 217).

Snaidero, Nicolas et al. (2014). "Myelin membrane wrapping of CNS axons by PI (3, 4, 5) P3-dependent polarized growth at the inner tongue". *Cell* 156.1, pp. 277–290 (pages 33, 34).

Solenov, Eugen et al. (2004). "Sevenfold-reduced osmotic water permeability in primary astrocyte cultures from AQP-4-deficient mice, measured by a fluorescence quenching method". *American Journal of Physiology-Cell Physiology* 286.2, pp. C426–C432 (page 53).

Somogyi, Peter et al. (1998). "Salient features of synaptic organisation in the cerebral cortex1". *Brain research reviews* 26.2-3, pp. 113–135 (page 31).

Song, L. et al. (2017). "On the Complexity of Learning Neural Networks". *ArXiv e-prints*. arXiv: 1707.04615 [cs.LG] (page 71).

Song, Sheng-Kwei et al. (2005a). "Demyelination increases radial diffusivity in corpus callosum of mouse brain". *Neuroimage* 26.1, pp. 132–140 (pages 169, 171).

– (2005b). "Demyelination increases radial diffusivity in corpus callosum of mouse brain." *Neuroimage* 26.1, pp. 132–40. DOI: 10.1016/j.neuroimage.2005.01.028 (page 170).

Song, Sheng-Kwei et al. (2003). "Diffusion tensor imaging detects and differentiates axon and myelin degeneration in mouse optic nerve after retinal ischemia". *Neuroimage* 20.3, pp. 1714–1722. DOI: 10.1016/j.neuroimage.2003.07.005 (pages 169–171).

Song, Sheng-Kwei et al. (2002). "Dysmyelination Revealed through MRI as Increased Radial (but Unchanged Axial) Diffusion of Water". *Neuroimage* 17.3, pp. 1429–1436 (pages 169, 183).

Sotelo, Constantino (2011). "Camillo Golgi and Santiago Ramon y Cajal: the anatomical organization of the cortex of the cerebellum. Can the neuron doctrine still support our actual knowledge on the cerebellar structural arrangement?" *Brain research reviews* 66.1, pp. 16–34 (page 14).

Sotiras, Aristeidis, Davatzikos, Christos, Paragios, Nikos (2013). "Deformable medical image registration: A survey". *IEEE transactions on medical imaging* 32.7, pp. 1153–1190 (pages 188, 190).

Springenberg, J. T. et al. (2014). "Striving for Simplicity: The All Convolutional Net". *ArXiv e-prints*. arXiv: 1412.6806 [cs.LG] (page 70).

Spruston, Nelson (2008). "Pyramidal neurons: dendritic structure and synaptic integration". *Nature Reviews Neuroscience* 9.3, p. 206 (page 19).

Srivastava, Nitish et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research* 15.1, pp. 1929–1958 (pages 73, 124).

Stanisz, Greg J et al. (1999). "Characterizing white matter with magnetization transfer and T2". *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 42.6, pp. 1128–1136 (pages 49, 183).

Stanisz, Greg J et al. (2005). "T1, T2 relaxation and magnetization transfer in tissue at 3T". *Magnetic resonance in medicine* 54.3, pp. 507–512 (page 49).

Stein, Wilfred (2012). *Transport and diffusion across cell membranes*. Elsevier (page 52).

Stejskal, Edward O, Tanner, John E (1965). "Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient". *The journal of chemical physics* 42.1, pp. 288–292 (pages 41, 43, 55).

Stevens, Charles F (1998). "Neuronal diversity: too many cell types for comfort?" *Current biology* 8.20, R708–R710 (page 18).

Steyerberg, Ewout W et al. (2010). "Assessing the performance of prediction models: a framework for some traditional and novel measures". *Epidemiology (Cambridge, Mass.)* 21.1, p. 128 (page 79).

Steyerberg, Ewout W et al. (2001). "Internal validation of predictive models: efficiency of some procedures for logistic regression analysis". *Journal of clinical epidemiology* 54.8, pp. 774–781 (pages 93, 94).

Stikov, Nikola et al. (2015a). "In vivo histology of the myelin g-ratio with magnetic resonance imaging". *Neuroimage* 118, pp. 397–405. DOI: 10.1016/j.neuroimage.2015.05.023 (page 184).

Stikov, Nikola et al. (2015b). "Quantitative analysis of the myelin g-ratio from electron microscopy images of the macaque corpus callosum". *Data in brief* 4, pp. 368–373 (page 50).

Stiles, Joan, Jernigan, Terry L (2010). "The basics of brain development". *Neuropsychology review* 20.4, pp. 327–348 (pages 24, 25, 30, 31).

Striedter, Georg F., Srinivasan, Shyam, Monuki, Edwin S. (2015). "Cortical Folding: When, Where, How, and Why?" *Annual Review of Neuroscience* 38.1, pp. 291–307. DOI: 10.1146/annurev-neuro-071714-034128 (page 15).

Studholme, Colin, Cardenas, Valerie (2004). "A template free approach to volumetric spatial normalization of brain anatomy". *Pattern Recognition Letters* 25.10, pp. 1191–1202 (page 192).

Sun, Chen et al. (2017). "Revisiting unreasonable effectiveness of data in deep learning era". *arXiv preprint arXiv:1707.02968* 1 (pages 64, 94).

Sun, Shu-Wei et al. (2006a). "Differential sensitivity of in vivo and ex vivo diffusion tensor imaging to evolving optic nerve injury in mice with retinal ischemia". en. *Neuroimage* 32.3, pp. 1195–1204 (page 170).

Sun, Shu-Wei et al. (2006b). "Noninvasive detection of cuprizone induced axonal damage and demyelination in the mouse corpus callosum." *Magnetic Resonance in Medicine* 55.2, pp. 302–8. DOI: 10.1002/mrm.20774 (pages 170, 171).

Sun, Wenjing, Dietrich, Dirk (2013). "Synaptic integration by NG2 cells". *Frontiers in cellular neuroscience* 7, p. 255 (page 19).

Sutton, Richard S, Barto, Andrew G (1998). *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge (page 64).

Suzuki, Kenji (2017). "Overview of deep learning in medical imaging". *Radiological Physics and Technology* 10.3, pp. 257–273. DOI: 10.1007/s12194-017-0406-5 (page 64).

Syková, Eva, Nicholson, Charles (2008). "Diffusion in brain extracellular space". *Physiological Reviews* 88.4, pp. 1277–1340 (page 175).

Szegedy, C. et al. (2013). "Intriguing properties of neural networks". *ArXiv e-prints*. arXiv: 1312.6199 [cs.CV] (pages 70, 156).

Szegedy, C. et al. (2015). "Rethinking the Inception Architecture for Computer Vision". *ArXiv e-prints*. arXiv: 1512.00567 [cs.CV] (pages 70, 71, 73).

Szegedy, Christian et al. (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning." *AAAI*. Vol. 4, p. 12 (page 126).

Takahashi, Emi et al. (2014). "Development of cerebellar connectivity in human fetal brains revealed by high angular resolution diffusion tractography". *Neuroimage* 96, pp. 326–333. DOI: 10.1016/j.neuroimage.2014.03.022 (page 232).

Talbott, Jason F et al. (2016). "Diffusion-Weighted Magnetic Resonance Imaging Characterization of White Matter Injury Produced by Axon-Sparing Demyelination and Severe Contusion Spinal Cord Injury in Rats". *Journal of Neurotrauma* (page 170).

Tax, Chantal M.W. et al. (2018). "The Dot... wherefore art thou? Search for the isotropic restricted diffusion compartment in the brain with spherical tensor encoding and strong gradients". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 26, p. 0253 (page 59).

Thiele, Thorvald Nicolai (1880). "Om Anvendelse af mindste Kvadraters Methode i nogle Tilfælde, hvor en Komplikation af visse Slags uensartede tilfældige Fejlkilder giver Fejlene en 'systematisk'Karakter". *Det Kongelige Danske Videnskabernes Selskabs Skrifter-Naturvidenskabelig og Mathematisk Afdeling*, pp. 381–408 (page 37).

Thirion, J-P (1998). "Image matching as a diffusion process: an analogy with Maxwell's demons". *Medical image analysis* 2.3, pp. 243–260 (page 190).

Thompson, Paul M et al. (1996). "Three-dimensional statistical analysis of sulcal variability in the human brain". *Journal of Neuroscience* 16.13, pp. 4261–4274 (page 199).

Tomassy, Giulio Srubek et al. (2014). "Distinct profiles of myelin distribution along single axons of pyramidal neurons in the neocortex". *Science* 344.6181, pp. 319–324 (page 31).

Torralba, Antonio, Efros, Alexei A (2011). "Unbiased look at dataset bias". *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, pp. 1521–1528 (page 152).

Torrey, Henry C (1956). "Bloch equations with diffusion terms". *Physical review* 104.3, p. 563 (page 43).

Tournier, J-Donald, Calamante, F, Connelly, A (2013). "Determination of the appropriate b value and number of gradient directions for high-angular-resolution diffusion-weighted imaging". *NMR in biomedicine*. DOI: 10.1002/nbm.3017/full (pages 182, 204, 219).

Tournier, J-Donald, Calamante, Fernando, Connelly, Alan (2007). "Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution". *Neuroimage* 35.4, pp. 1459–1472 (pages 56, 58, 60).

Tournier, J-Donald et al. (2004). "Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution." *Neuroimage* 23.3, pp. 1176–85. DOI: 10.1016/j.neuroimage.2004.07.037 (pages 58, 60).

Tournier, J-Donald et al. (2008). "Resolving crossing fibres using constrained spherical deconvolution: validation using diffusion-weighted imaging phantom data". *Neuroimage* 42.2, pp. 617–625 (page 191).

Tournier, Jacques-Donald et al. (2015a). "Data-driven optimisation of multi-shell HARDI". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 23 (pages 115, 204, 217).

Tournier, Jacques-Donald et al. (2015b). "Optimisation of single-shell HARDI for neonatal imaging". *Proc. Intl. Soc. Mag. Reson. Med.* Vol. 23 (pages 115, 204, 236).

Trouard, Theodore P et al. (1996). "Analysis and comparison of motion-correction techniques in diffusion-weighted imaging". *Journal of Magnetic Resonance Imaging* 6.6, pp. 925–935 (page 47).

Tsunoda, Kazushige et al. (2001). "Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns". *Nature neuroscience* 4.8, pp. 832–838 (page 23).

Tuch, David S et al. (2002a). "High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity". *Magnetic resonance in medicine* 48.4, pp. 577–582 (page 55).

Tuch, David Solomon et al. (2002b). "Diffusion MRI of complex tissue structure". PhD thesis. Massachusetts Institute of Technology (page 55).

Tustison, Nicholas J et al. (2010). "N4ITK: improved N3 bias correction". en. *IEEE transactions on medical imaging* 29.6, pp. 1310–1320. DOI: 10.1109/TMI.2010.2046908 (pages 196, 204, 217).

Uğurbil, Kamil et al. (2013). "Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project". *Neuroimage* 80, pp. 80–104 (page 196).

Vaillant, Marc et al. (2004). "Statistics on diffeomorphisms via tangent space representations". *Neuroimage* 23, S161–S169 (page 193).

Valiente, Manuel, Marín, Oscar (2010). "Neuronal migration mechanisms in development and disease". *Current opinion in neurobiology* 20.1, pp. 68–78 (page 24).

Valle, Eduardo et al. (2017). "Data, Depth, and Design: Learning Reliable Models for Melanoma Screening". *arXiv preprint arXiv:1711.00441* (pages 68, 94, 102, 130, 149).

Van Kooij, Britt JM et al. (2011). "Fiber tracking at term displays gender differences regarding cognitive and motor outcome at 2 years of age in preterm infants". *Pediatric research* 70.6, p. 626 (page 99).

Vanderhaeghen, Pierre, Cheng, Hwai-Jong (2010). "Guidance molecules in axon pruning and cell death". *Cold Spring Harbor perspectives in biology* 2.6, a001859 (page 30).

Varentsova, Anna, Zhang, Shengwei, Arfanakis, Konstantinos (2014). "Development of a high angular resolution diffusion imaging human brain template". *Neuroimage* 91, pp. 177–186 (page 187).

Veraart, Jelle et al. (2016). "Denoising of diffusion MRI using random matrix theory". en. *Neuroimage* 142, pp. 394–406. DOI: 10.1016/j.neuroimage.2016.08.016 (pages 204, 217).

Vercauteren, Tom et al. (2009). "Diffeomorphic demons: Efficient non-parametric image registration". *Neuroimage* 45.1, S61–S72 (page 190).

Verhey, Leonard H., Shroff, Manohar, Banwell, Brenda (2013). "Pediatric Multiple Sclerosis. Pathobiological, Clinical, and Magnetic Resonance Imaging Features". *Neuroimaging Clinics of North America* 23.2, pp. 227–243. DOI: 10.1016/j.nic.2012.12.004 (page 169).

Verkman, AS et al. (1996). "Water transport across mammalian cell membranes". *American Journal of Physiology-Cell Physiology* 270.1, pp. C12–C30 (page 49).

Viergever, Max A et al. (2016). "A survey of medical image registration–under review". *Medical image analysis* 33, pp. 140–144 (page 187).

Vinyals, O. et al. (2016). "Matching Networks for One Shot Learning". *ArXiv e-prints*. arXiv: 1606.04080 [cs.LG] (page 76).

Vlahos, L. et al. (2008). "Normal and Anomalous Diffusion: A Tutorial". *ArXiv e-prints*. arXiv: 0805.0419 [nlin.CD] (page 37).

Volpe, Joseph J (2001). "Neurobiology of periventricular leukomalacia in the premature infant". *Pediatric research* 50.5, p. 553 (page 33).

Volpe, Joseph J. (2008). *Neurology of the Newborn, Volume 899*. Elsevier Health Sciences, p. 1094 (pages 16, 17, 21, 24, 28).

Vos, Sjoerd B et al. (2011). "Partial volume effect as a hidden covariate in DTI analyses". *Neuroimage* 55.4, pp. 1566–1576 (page 171).

Wake, Hiroaki, Lee, Philip R, Fields, R Douglas (2011). "Control of local protein synthesis and initial events in myelination by action potentials". *Science* 333.6049, pp. 1647–1651 (page 34).

Wang, Yong et al. (2015). "Differentiation and quantification of inflammation, demyelination and axon injury or loss in multiple sclerosis". *Brain* 138.Pt 5, pp. 1223–1238 (page 184).

Wang, Yong et al. (2011). "Quantification of increased cellularity during inflammatory demyelination". en. *Brain* 134.12, pp. 3590–3601 (page 171).

Wansapura, Janaka P et al. (1999). "NMR relaxation times in the human brain at 3.0 tesla". *Journal of magnetic resonance imaging* 9.4, pp. 531–538 (page 48).

Warach, S et al. (1992). "Fast magnetic resonance diffusion-weighted imaging of acute human stroke". *Neurology* 42.9, pp. 1717–1717 (page 53).

Watkins, Trent A et al. (2008). "Distinct stages of myelination regulated by γ-secretase and astrocytes in a rapidly myelinating CNS coculture system". *Neuron* 60.4, pp. 555–569 (pages 33, 34).

Waxman, S G, Bennett, MVL (1972). "Relative conduction velocities of small myelinated and non-myelinated fibres in the central nervous system". *Nature* (page 183).

Waxman, SG, Pappas, GD, Bennett, MVL (1972). "Morphological correlates of functional differentiation of nodes of Ranvier along single fibers in the neurogenic electric organ of the knife fish Sternarchus". *The Journal of cell biology* 53.1, pp. 210–224 (page 31).

Wedeen, Van J et al. (2005). "Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging". *Magnetic resonance in medicine* 54.6, pp. 1377–1386 (page 53).

Wehrens, Ron, Putter, Hein, Buydens, Lutgarde MC (2000). "The bootstrap: a tutorial". *Chemometrics and intelligent laboratory systems* 54.1, pp. 35–52 (page 93).

Weiss, Gary M, Provost, Foster (2003). "Learning when training data are costly: The effect of class distribution on tree induction". *Journal of Artificial Intelligence Research* 19, pp. 315–354 (page 79).

Wheeler Kingshott, Claudia A M, Cercignani, Mara (2009). "About "axial" and "radial" diffusivities". *Magnetic Resonance in Medicine* 61.5, pp. 1255–1260 (pages 170, 236).

Wheeler-Kingshott, Claudia A M et al. (2012). "A new approach to structural integrity assessment based on axial and radial diffusivities". *Functional Neurology* 27.2, pp. 85–90 (page 169).

White, LE et al. (1997). "Structure of the human sensorimotor system. I: Morphology and cytoarchitecture of the central sulcus." *Cerebral cortex (New York, NY: 1991)* 7.1, pp. 18–30 (page 199).

White, Nathan S et al. (2014). "Diffusion-weighted imaging in cancer: physical foundations and applications of restriction spectrum imaging". *Cancer research* 74.17, pp. 4638–4652 (page 49).

Wilhelm, Michael J et al. (2012). "Direct magnetic resonance detection of myelin and prospects for quantitative imaging of myelin density." *Proceedings of the National Academy of Sciences of the United States of America* 109.24, pp. 9605–10. DOI: 10.1073/pnas.1115107109 (page 49).

Wilson, D Randall, Martinez, Tony R (2003). "The general inefficiency of batch training for gradient descent learning". *Neural Networks* 16.10, pp. 1429–1451 (page 73).

Wimberger, D.M. et al. (1995). "Identification of "Premyelination" by diffusion-weighted MRI". *Journal of Computer Assisted Tomography* 19.1, pp. 28–33. DOI: 10.1097/00004728-199501000-00005 (page 230).

Wolpert, David H (1995). "Off-training set error and a priori distinctions between learning algorithms". *Sante Fe Institute, Santa Fe, NM, USA, Tech. Rep. SFI-TR*, pp. 95–01 (page 92).

Wong, S.C. et al. (2016). "Understanding data augmentation for classification: when to warp?" *ArXiv e-prints*. arXiv: 1609.08764 [cs.CV] (page 74).

Xiao, H., Rasul, K., Vollgraf, R. (2017). "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms". *ArXiv e-prints*. arXiv: 1708.07747 [cs.LG] (page 104).

Xie, Mingqiang et al. (2010). "Rostrocaudal analysis of corpus callosum demyelination and axon damage across disease stages refines diffusion tensor imaging correlations with pathological features". en. *Journal of Neuropathology and Experimental Neurology* 69.7, pp. 704–716 (page 170).

Xu, Tianyou et al. (2015). "The role of myelin geometry on magnetic susceptibility-driven frequency shifts: toward realistic geometries". *ISMRM 23rd Annual Meeting & Exhibition* (page 183).

Yablonskiy, Dmitriy A, Sukstanskii, Alexander L (2010). "Theoretical models of the diffusion weighted MR signal". *NMR in Biomedicine* 23.7, pp. 661–681 (pages 42, 53, 54).

Yakovlev, Paul I., Lecours, Andre-Roch (1967). "The myelogenetic cycles of regional maturation of the brain". *Regional development of the brain in early life*, pp. 3–70 (pages 35, 230, 231).

Yang, Donghan M et al. (2017). "Intracellular water preexchange lifetime in neurons and astrocytes". *Magnetic resonance in medicine* (page 52).

Yaniv Assaf et al. (2004). "New modeling and experimental framework to characterize hindered and restricted water diffusion in brain white matter". en. *Magnetic Resonance in Medicine* 52.5, pp. 965–978. DOI: 10.1002/mrm.20274 (page 174).

Yanovsky, Igor et al. (2008). "Asymmetric and symmetric unbiased image registration: statistical assessment of performance". *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, pp. 1–8 (page 190).

Yeh, Fang-Cheng, Tseng, Wen-Yih Isaac (2011). "NTU-90: a high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction". *Neuroimage* 58.1, pp. 91–99 (page 187).

Yeo, BT Thomas et al. (2009). "DT-REFinD: Diffusion tensor registration with exact finite-strain differential". *IEEE transactions on medical imaging* 28.12, pp. 1914–1928 (page 191).

Yoshida, Shoko et al. (2013). "Diffusion tensor imaging of normal brain development". *Pediatric radiology* 43.1, pp. 15–27 (page 10).

Yoshiura, T et al. (2001). "Highly diffusion-sensitized MRI of brain: Dissociation of gray and white matter". *Magnetic resonance* (page 182).

Yosinski, J. et al. (2014). "How transferable are features in deep neural networks?" *ArXiv e-prints*. arXiv: 1411.1792 [cs.LG] (page 101).

Younes, Laurent (2007). "Jacobi fields in groups of diffeomorphisms and applications". *Quarterly of applied mathematics*, pp. 113–134 (page 193).

Young, Kaylene M et al. (2013). "Oligodendrocyte dynamics in the healthy adult CNS: evidence for myelin remodeling". *Neuron* 77.5, pp. 873–885 (page 31).

Yushkevich, Paul A et al. (2008). "Structure-specific statistical mapping of white matter tracts". *Neuroimage* 41.2, pp. 448–461 (page 187).

Zagoruyko, S., Komodakis, N. (2016). "Wide Residual Networks". *ArXiv e-prints*. arXiv: 1605.07146 [cs.CV] (page 71).

Zanin, Emilie et al. (2011). "White matter maturation of normal human fetal brain. An in vivo diffusion tensor tractography study." *Brain Behav.* 1.2, pp. 95–108. DOI: 10.1002/brb3.17 (page 230).

Zaqout, Sami, Kaindl, Angela M (2016). "Golgi-cox staining step by step". *Frontiers in neuroanatomy* 10, p. 38 (page 27).

Zeiler, M. D, Fergus, R. (2013). "Visualizing and Understanding Convolutional Networks". *ArXiv e-prints*. arXiv: 1311.2901 [cs.CV] (pages 70, 157).

Zeng, Hongkui, Sanes, Joshua R (2017). "Neuronal cell-type classification: challenges, opportunities and the path forward". *Nature Reviews Neuroscience* 18.9, p. 530 (page 18).

Zhan, Wang, Yang, Yihong (2006). "How accurately can the diffusion profiles indicate multiple fiber orientations? A study on general fiber crossings in diffusion MRI". *Journal of Magnetic Resonance* 183.2, pp. 193–202 (page 191).

Zhang, C. et al. (2016). "Understanding deep learning requires rethinking generalization". *ArXiv e-prints*. arXiv: 1611.03530 [cs.LG] (pages 72, 73).

Zhang, Hui et al. (2006). "Deformable registration of diffusion tensor MR images with explicit orientation optimization". *Medical image analysis* 10.5, pp. 764–785 (page 191).

Zhang, Hui et al. (2012). "NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain". *Neuroimage* 61.4, pp. 1000–1016 (pages 56, 57).

Zhong, Z. et al. (2017). "Random Erasing Data Augmentation". *ArXiv e-prints*. arXiv: 1708.04896 [cs.CV] (page 74).

Zhou, Zhenyu et al. (2011). "Automated artifact detection and removal for improved tensor estimation in motion-corrupted DTI data sets using the combination of local binary

patterns and 2D partial least squares". *Magnetic resonance imaging* 29.2, pp. 230–242 (page 100).

Zhu, Kangrong et al. (2016). "Hybrid-Space SENSE Reconstruction for Simultaneous Multi-Slice MRI". *IEEE Transactions on Medical Imaging* 35.8, pp. 1824–1836. DOI: 10.1109/TMI.2016.2531635 (pages 115, 217).

Zhu, Xiangxin et al. (2012). "Do We Need More Training Data or Better Models for Object Detection?." *BMVC*. Vol. 3, p. 5 (pages 101, 108).

Zitova, Barbara, Flusser, Jan (2003). "Image registration methods: a survey". *Image and vision computing* 21.11, pp. 977–1000 (page 188).

Zöllei, Lilla et al. (2005). "Efficient population registration of 3D data". *International Workshop on Computer Vision for Biomedical Image Applications*. Springer, pp. 291–301 (page 192).

Zoph, Barret, Le, Quoc V (2016). "Neural architecture search with reinforcement learning". *arXiv preprint arXiv:1611.01578* (page 152).

Zoph, Barret et al. (2017). "Learning transferable architectures for scalable image recognition". *arXiv preprint arXiv:1707.07012* (page 152).

Zwiers, Marcel P (2010). "Patching cardiac and head motion artefacts in diffusion-weighted images". *Neuroimage* 53.2, pp. 565–575 (page 98).

List of Figures

List of Tables